Introduction of Statistics

Book · December 2018

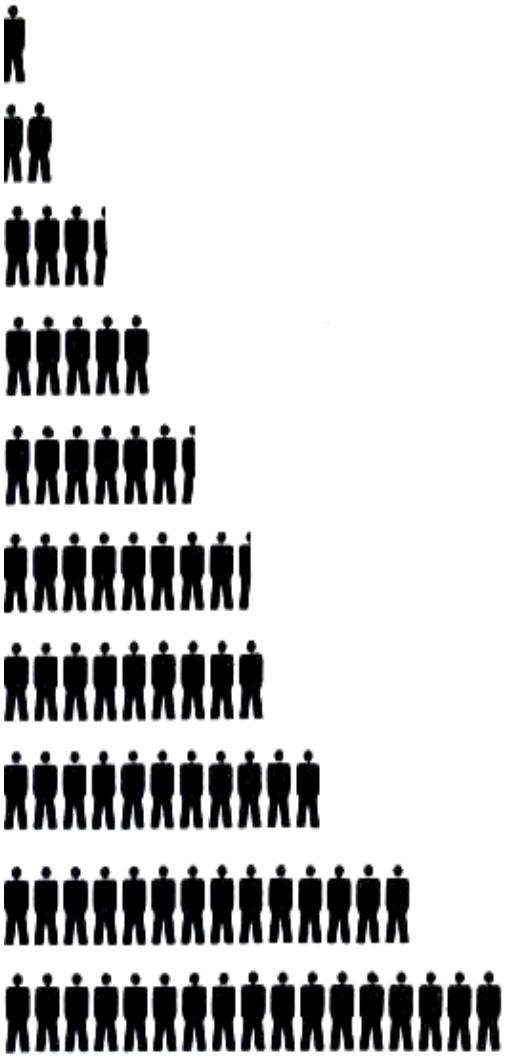
CITATIONS
0

READS
51,975

1 author:

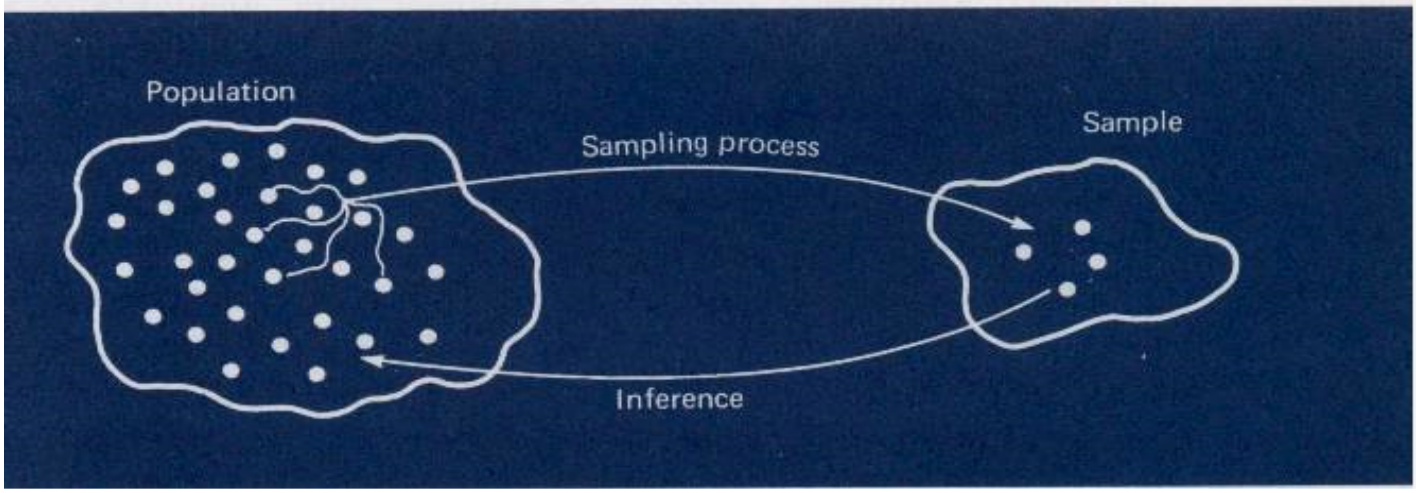
Z. A. Al-Hemyari
Independent Researcher
113 PUBLICATIONS **527** CITATIONS

SEE PROFILE



UNIVERSITY OF NIZWA

Introduction to Statistics



Prof. Dr. Zuhair A. Al-Hemyari

University of Nizwa

College of Arts and Sciences

Department of

Mathematical and Physical Sciences

Section : Statistics

STAT101: Introduction to Statistics

Lecture/Tutorial (3:2) 4CR

Brief description of what statistics is all about. Common areas of use. Types job opportunities. Sources of data : Internal data and external data .Types of data: Nominal, ordinal, interval and ratio. Statistical data terminology : Population, population characteristics, sample ,population census , sampling and non-sampling errors,… and variables. Obtaining data. Descriptive statistics: Summarizing data. One and two way frequency tables and how to make them. Scatter plots. Measures of central tendency, mean, median, mode (grouped and non-grouped data) . Measures of variability, standard deviation, range, skewness measure (grouped and non-grouped data).Quartiles and percentiles. Definitions of probability. Random variables and their distributions . Simple uses/application of binomial and normal distributions. Simple indices and rates . Scatter plots and simple linear regression.

STAT105: Statistics for Engineers

(Lecture/Tutorial (2:2) 3CR

This course is an introduction to the basic concepts of probability and statistics. The examples and exercises strongly emphasize engineering applications.

Introduction to Statistics

STAT101

Instructor :

Prof. Dr. Zuhair A. Al-Hemyari

Contents

3.7.1 Geometric mean
3.7.2 Quartiles and Percentiles

Tutorials 3.1 & 3.2

- # Measures of Variation

4.1 Introduction
4.2 The range
4.3 The standard deviation
4.4 Measure of relative variation
4.5 Measure of skewness
Tutorial 4

- # Introduction to Probability & Random Variables

5.1 Introduction
5.2 The sample and event spaces
5.3 Computing probabilities from the sample space
5.4 Permutations, combinations, and other counting rules
5.5 Random variable
5.6 Probability mass function
5.7 Probability density function
Tutorials 5.1 & 5.2

- **Binomial & Normal distributions**

6.1 Probability function
6.2 The binomial distribution
6.3 The normal distribution
Tutorial 6

- **Regression Analysis**

7.1 Introduction
7.2 Relationships between variables
7.3 Simple linear regression model
7.4 Fitting of a simple linear regression model
Tutorial 7

Probability Tables
1. Binomial cumulative distribution function.
2. Standard Normal distribution function

References

Introduction

to

Statistics

1

Introduction

1.1 Examples of Statistical Problems

The word "***statistics***" conveys a variety of meanings to people, many of which are inaccurate or, at the very least, misleading. To some, the word suggests only a plethora of mind-boggling tables, charts, and figures. Other people consider statistics to be an imposing form of mathematics. The use of the word certainly had an inauspicious beginning, as might be suspected from a cursory study of the word, for it was originally a term used to denote a collection of figures, graphs and the like which contained useful information for the state (primarily budget information such as taxation figures).

Used in the context of its original meaning, statistics generally refers to information about an activity or a process that is expressed in numbers listed in tables or illustrated in figures. But, since its early connotation, statistics has grown to encompass a larger role than presenting us with charts, graphs, and tables or figures. In a modern setting, ***statistics*** refers to the science of collecting, presenting, and analyzing numerical data. A ***statistician*** is a person who engages in one or more of the following tasks:

(1) the clerical activities of tabulating, summarizing, and displaying statistical data.

(2) analyzing data by using statistical methods, usually for the purposes of decision making.

(3) advancing the science of statistics by developing new and better analysis methods.

The levels of expertise required by statisticians ranges from mastering simple clerical operations with data to advanced training in applied mathematics, and statisticians are needed at all levels.

The use of statistics has permeated almost every facet of our lives. The daily newspapers and the televised news reports supply us with numerous summaries of data such as stock market reports, financial summaries, and crime statistics-and with the results of statistical analyses-weather forecasts, political election outcome predictions, and so on.

Governments, businesses, and individuals collect statistical data required to carry out their activities efficiently and effectively. The rate at which statistical data are being collected is staggering and is primarily due to the realization that better decisions are possible with more information and, perhaps more importantly, to technological advances that have enabled the efficient collection and analysis of large bodies of data.

The most important technological advance in this area has, of course, been the development of the electronic digital computer. Statistical concepts and methods, and the use of computers in statistical analyses, have affected virtually all disciplines biology, physics, engineering, economics, sociology, psychology, business, and others. In business and economics, the development and application of statistical methods have led to greater production efficiency, to better forecasting techniques, and to better management practices. It is becoming increasingly apparent that some knowledge of statistics and computers is essential for careers in economics, business, administration, and many other fields as well. To gain an appreciation for the breadth of applications of statistics to business and economic problems in particular, let us consider three examples.

Example 1.1

In operations management, a primary concern is controlling the quality of the items being produced. If the product is a transistor radio battery, for example, we may be concerned with the longevity of the batteries. Suppose it is desired that at least 95 percent of the batteries last through at least 20 hours of continuous use. The actual percentage of batteries lasting m ore than 20 hours could be determined by inserting each and every battery produced into a transistor radio and recording its time to failure, but then there would be no batteries to sell. Rather, a manager may wisely decide in a day's production to pull every 100th battery off the production line, insert the sampled batteries in electrical test circuits and record their times to failure. The percentage of these batteries lasting through more than 20 hours of continuous use could be used to estimate the percentage of all batteries produced during that day which will last more than 20 hours. Moreover, if this estimated percentage drops much below 95 percent (say to 80 percent), the manager may wish to stop the production line until he can determine why the percentage of bad batteries appears to be greater than the tolerated 5 percent. The manager is using a percentage statistic computed from a sample of all batteries produced to arrive at a decision regarding the quality of the set of all batteries produced on a given day.This example portrays a common phenomenon in quality control: destructive sampling. It is impossible to test the quality (longevity) of each battery produced because the test for longevity will ordinarily involve its destruction.

The manager has little recourse but to sacrifice a small number of batteries (the *sample)* in order to gain information about the entire set of batteries comprising the daily production (the population*).*

Example 1.2

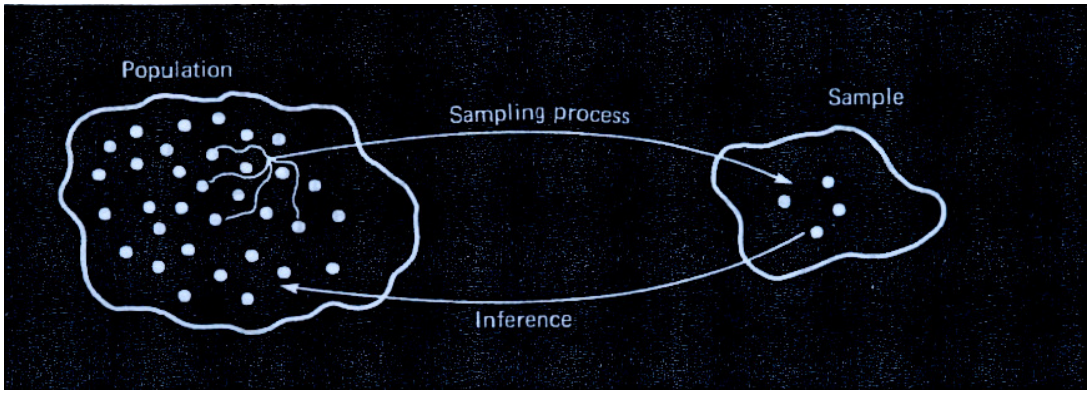
Determining the saleability of a new product is a constant problem posed to many marketing research groups. In order to determine whether a new kitchen-ware product will sell, the marketers might conduct a house-to house survey of 1,000 households selected randomly in the product target areas, during which they present the product to the housewife for evaluation. The percentage of the housewives willing to buy the product at its listed price, together with other information obtained from the interviews could be used to decide whether or not the new item should undergo full-scale production.

Example 1.3

Politicians and their supporters aelection as the campaign heads towards final balloting. By sampling 1,000 regire immensely interested in knowing their prospects of winning an stered voters prior to the election, the percentage who claim they will vote for a given candidate may be used to estimate the percentage of the votes the candidate will receive hi the election. The estimated percentage could be used to decide, for example, whether a greater campaign effort (more money) is required to assure the candidate's election. There are many more examples in business and other areas which might be cited, but the above five should indicate the many ways in which statistics can be employed. In the first two examples, statistics is used to *describe* large bodies of data. In this application, the word "statistic" is being used to describe a specific numerical quantity such as an average or a total, and the collection of statistics is used to summarize or condense a large set of numbers. These compiled statistics may, in turn, be used to assist in decision making. In the last three examples, statistics may be interpreted in a much broader sense; namely, the process of drawing conclusions about an entire populatio*n* or collection of things based upon a sample*,* a subset of the population or collection.

Most students probably view statistics in the context of the first example above; that is, as tables of figures, charts, and graphs (batting averages, pie charts illustrating the sources of government revenue, and so on). This concept is called descriptive statistics and was at one time the principal use of statistics in business. Currently, there is an increasing interest in the methods and uses of inferential statistics-the process of drawing inferences about the whole (the population) from a subset of it (the sample), as exemplified in Examples 1.1-1.3. Schematically, the process of drawing inferences about an unknown population numerical quantity (the proportion of defectives in a production lot, mean incomes of a class of laborers, etc..) is illustrated in figure 1.1. Units are selected from the population to form the sample which in turn is used to draw inferences about the population characteristic of interest. Much of this text is devoted to the study of statistical inference. In subsequent sections of this chapter, we will focus attention on the sources of data, methods of obtaining data, and data measurement considerations.

Figure 1.1 The Statistical Inference Process



1.2 Sources of Data

In most instances, businesses use ***internal data***; that is: data arising from bookkeeping practices, standard operating, business procedures, or planned experiments by *research divisions* within the company. Examples are profit and loss statements, employee salary information, production data and economic forecasts. Occasionally, it may be necessary or desirable to use sources of ***external data***. By external,

We mean:
sources of data outside the firm. ***External data*** may be of two types: ***primary data*** and ***secondary data***. By *primary data* , we mean data obtained from the organization which originally collected them. An example is the population data collected by and available from the US. Bureau of the Census. *Secondary data* come from a source other than the one which originally collected them.

Ordinarily, if external data must be used, it is recommended that primary data be sought out since it will not have undergone any "refining" by the secondary source.

In the election survey (Example 1.3) in Section l. l, the Statistical Abstract provide numerous tables of both primary and secondary data, such as past voting records in districts and numbers of registered Democrats and Republicans, which may supply important information in conjunction with the internal sampled data on estimating a candidate's probability of being elected.
There are many excellent sources of published (primary and secondary) data which have been compiled by the state and government, by business and economic associations, and by commercial sources (periodicals).
Some examples are: The Statistical Abstract of the United States (published annually by the Bureau of the Census), Survey of Current Business (published by the Department of Commerce), Monthly Labor Review (published by the Bureau of Labor Statistics), Harvard Business Review (periodical), Business Week (periodical), The Wall Street Journal (periodical), Dun's Statistical Review (periodical) and The Journal of Management Science (association journal). Additional sources of external data, available in most reference libraries, are The Economic Almanac, Federal Reserve Bulletin, Life Insurance Fact Book, International Financial Statistics, and Business Conditions Digest.
Caution must always be exercised in using external sources of data, particularly secondary sources, as they may contain errors in transcription from the primary source. When external data are used, the conditions under which the data were collected and summarized must be determined to assure that they are relevant for the intended use. This determination usually requires identifying and locating the primary source,

which typically will discuss any restrictions placed on the data due to the process of their collection. Thus, while secondary sources of data are convenient, it usually is prudent to seek out and use primary sources of external data.

1.3 Statistical Data Terminology

When statistical data are collected and analyzed, it is usually in the context of populations and their characteristics.

Definition 1.1

Population and population characteristic

A *population* is the totality of units under study. A *population characteristic* is an attribute of a population unit**.**

We may be interested, for example, in the salaries of workers in a particular industry. If so, the population is the totality of these workers and the characteristic of interest is each worker's salary. In collecting the salary data, we may be interested in other population characteristics as well, including sex, age, educational level, and other information. In general, a population unit may have one or more characteristics of interest in a particular study.

As another illustration of a population, a firm may be interested in the proportion of defective units of a certain product that it has produced in a large lot stored in a warehouse. The population is the totality of units in the warehouse and the characteristic is the acceptability of each unit of the product-it is either defective or non-defective .

A population may be either entirely inspected or partially inspected. When the data are produced by measuring the population characteristic for each and every unit in the population, we say a census of the population has been taken.

Definition 1.2

A population census

A *population census* is the evaluation of each and every unit in the population under study.

In some situations, it is possible to take a complete census of the population. This rarely occurs in business unless the population size is very small, due to cost and time considerations. A census of the US. population is undertaken every ten years and it is truly a Herculean effort, subsidized, naturally, by the taxpayers. The US. census produces a wealth of data of considerable importance to the federal government and to firms and institutions, many of whom view the census as an important source of external data.

In most instances, it is not possible to take a *census* of a population. It may be too costly, too time consuming, or the evaluation process may destroy the population unit as in Example 1.1.

Definition 1.3

A sample

A *sample* is a part of a population in which the population characteristic is studied so that inferences may be made from the sample study about the entire population.

A classic example of a situation in which samples must be used rather than a census taken is destructive sampling in which the process of evaluating a unit of the population destroys or irrevocably damages that unit.

Example 1.4

Suppose a tire manufacturer wishes to claim its new radial tire will last 40,000 miles or more. To support this claim, a sample of all tires produced (the population) is selected for testing to determine how many miles the tires will last. Since testing destroys the tires, a complete census of the population is impossible.

The *advantages of sampling* to taking a census are rather obvious. A sample is less expensive than a census, it can produce data more quickly, and the data are often more reliable because more time can be spent studying each sampled unit. But there clearly is a price to be paid as well. By looking at only a portion of the population, we are subject to *errors* because the sample

may not be representative of the whole population.

Definition 1.4
Sampling error

Sampling error is the difference between the result of studying a sample and infering a result about the population, and the result of a census of the whole population.

Example 1.5

As an illustration of sampling error, suppose we are interested in the average salary of unionized workers in a specific industry and we know from union membership lists that there are presently 1,000 workers in this industry. Had we taken a census, we may have found that the average salary is, let us say, $25,000. Based upon a sample of 100 workers, we might find that the average salary is $27,200. The difference between the two figures- $2,200 is the sampling error.

Errors in acquiring and tabulating statistical data can arise in other ways as well, and these errors are called nonsampling errors.

Definition 1.5
Nonsampling error

Nonsampling errors are errors that occur in acquiring, recording, or tabulating statistical data that cannot be ascribed to sampling error. They may arise in either a census or a sample.

Nonsampling errors are usually more difficult to control and detect than sampling errors.

Example 1.6

Suppose we are acquiring data on the 1,000 unionized workers mentioned above. If we approached a particular worker and asked for his or her income, we could be lied to-a troublesome and frequent source of nonsampling error when a sensitive question is asked directly of a person. (What is your grade point average?) In some instances, a person may give a false response out of ignorance rather than by design. Another source of nonsampling error is in recording the data. A"7" may be written as a "9," the decimal point may be incorrectly placed, and so on. Errors may also occur in tabulating the data-keypunching errors in preparing computer cards and typing errors in

transcribing data, for instance. It is always necessary to carefully edit data to minimize the chance of nonsampling errors adversely affecting the statistical analysis of the data.
The identification of the units in a population under study can often be a surprisingly difficult task. We refer to a listing of population units as a frame.

Definition 1.6
Population frame
The listing of all units in the population under study is called the *population frame.*

Example 1.7

If the population is a production lot of units stored in a warehouse, production records will give us a listing of the serial numbers of the units from which each unit may be identified.

Example 1.8

If the population is the 1,000 unionized workers in a specific industry, union membership records may serve as a frame.

But, what about a frame for all persons who will vote in a particular election? A listing of registered voters is not appropriate, because in many elections less than 50 percent of the registered voters actually vote. Some classic errors have been made in identifying the frame.

1.4 The acquisition of Data: Surveys and Experiments

When internal or external data are not readily available or are incomplete in a study attempting to answer questions about a population, a survey or an experiment may be conducted to provide the required information.

Definition 1.7
Statistical survey
A *survey* is a process of collecting data from existing population units, with no particular control over factors that may affect the population characteristics of interest in the study.

Most of us are very familiar with surveys. As students, we are asked about our opinions regarding dining hall food, impending tuition hikes, teaching effectiveness and so on. Filling out survey questionnaires or answering an interviewer's questions has become a routine occurrence in most of our lives.

Example 1.9

suppose we are interested in acquiring data on the salaries of 1,000 unionized workers in a specific industry. The population characteristic "salary" may be affected by a host of factors-age , race, sex, educational level, etc.. As we elicit a particular worker's salary, we have no control over educational level, age, and so on these are existing attributes of the worker.

In contrast to a survey is a statistical experiment in which we do exercise control over factors that may affect the population characteristics of interest.

Definition 1.8

Statistical experiment

An *experiment* is a process of collecting data about population characteristics when control is exercised over some or all factors that may affect the characteristics of interest in the study.

Example 1.10

We may be interested, in the yield of a chemical process that is affected by temperature and pressure. A variety of settings for temperature and pressure could be selected, and the chemical process run for each setting to determine the yield. In this way, the joint effect of temperature and pressure on yield is studied in a controlled manner.

Example 1.11

In management, we may be interested in the effects of a training program on the first year performance of new employees. A set of new employees may be split into two groups such that both groups are approximately alike in terms of age, sex, education, and other factors. The training program could be administered to one group and not to the other (the control group). At the end of the first year, performance characteristics could be measured to assess the effects of the training program, accounting for factors other than the training program that may affect performance.

Experiments almost always provide better information than do surveys, but both are extremely important and useful tools for acquiring data. Though an experiment should be preferred to a survey, much of the data used in statistical analyses in business and economics are survey data. There are a number of reasons for this. First, most internal and external data are collected by surveys. Second, it is not always possible to conduct an experiment to acquire the needed information. An interesting example of this is the effect of smoking on health. Virtually all data on the relationship between smoking and health are survey data; other factors that may affect health, such as age, race, sex, and physiological properties, are not in the control of those collecting the data. To run an experiment in this case would involve controlling persons' lives. Some people in the experiment would be required to smoke while others would not. It is neither feasible nor desirable to approach the acquisition of the data for the study of the relationship between smoking and health in this way.

The *planning* of a *survey* or an *experiment* is essential to ensure that the resulting information will be useful. A good plan usually involves the following steps (these steps are applied to the quality control problem in Example 1.1):

1. A clear and detailed statement of the problem.

The statement of the problem should clearly indicate that we are interested in determining whether or not the percentage of good batteries (those lasting through 20 hours) exceeds a specified percentage (95 percent). The population is comprised of all batteries produced during a chosen period of time (a day or a week) and the characteristic in the population of interest is the number of hours the battery will continuously operate before failure.

2. A decision to survey or to experiment.
In this problem, it is possible to answer the question about the population by experimentation. We may test each selected battery under the set of conditions in which it was designed to operate.

3. A decision to take a census or a sample.
This is a case of destructive sampling. In determining the proportion of batteries that will last 20 or more hours, the tested batteries are "spent." We must, therefore, take a sample of batteries.

4. Designing the survey or experiment.
The experiment must be designed so that we isolate the characteristic of interest-the lifetime of the battery. Test circuits must be constructed and carefully monitored when the selected batteries are inserted for testing.

5. Collecting and analyzing the data.
For each battery, the time to failure is recorded and the proportion of batteries lasting 20 hours or more is calculated.

6. Reaching conclusions about the population characteristics.
The sample proportion of batteries surviving 20 or more hours is used as an estimate of the population proportion that survive 20 or more hours.

7. Reporting the results.
The report should include a thorough description of the problem, the sampling design, the testing method and the inferences. Sufficient monies should be allocated for a competent writing of the report. Indeed, many companies employ technical writers to put into "laymen's" words the experimental results.

The manner in which a sample is drawn, the methods of analyzing statistical data , and the kinds of inferences that may be drawn from the analysis (steps 4, 5 , and 6) are major topics in this text. It is important not to minimize the other steps , particularly steps I and 7. A clear and detailed statement of the problem is essential in planning a survey or an experiment. And the best analysis of survey or experimental results is meaningless unless the analysis can be accurately and understandably reported.

1.5 Obtaining Data

Once it has been determined that a survey or experiment is required, there are a variety of methods that may be employed. The most difficult problems arise when gathering information from people in surveys and the methods most relevant to this situation will be emphasized.

1.5.1 Self -Enumeration

Self-enumeration is probably the most common method of acquiring data from people in a survey or in an experiment. Questionnaires are usually distributed to selected individuals by mail,

although the distribution mechanism depends to a large extent on the purpose and nature of the questionnaire. For example, if the purpose of the questionnaire is to survey the attitudes of those using public transportation, the questionnaire may be distributed to persons while commuting to and from work on buses, subways, and trains.

The use of *questionnaires suffers* from two serious drawbacks. First, if the respondent has difficulty in interpreting the questions, no one is available for assistance. If this situation arises, the information received may contain a high degree of nonsampling error or the respondent may become frustrated and not bother completing or returning the questionnaire. Further, if a questionnaire is mailed to a household, it is often not clear who in the household responded to it. Second, questionnaires have typically an extremely poor response rate. It is not uncommon to have less than 30 percent returned on the first mailing of a questionnaire. The principal advantage of a *questionnaire* is the low cost relative to the other means of obtaining information. Most mail questionnaires *may* be bulk mailed at a reasonable rate. But it is almost always necessary to contact nonrespondents to the first mailing by subsequent mailings, telephone calls, or personal interviews, and these costs must be planned for in a well-designed self-enumeration survey or experiment. In most instances, those who do respond to the first mailing of a questionnaire are not representative of the entire population. To use only their responses would tend to bias the analytical results. Some self-enumeration questionnaires do enjoy high initial response rates. Examples are questions asked on warranty cards that must be returned to the manufacturer for warranty coverage of a new product .

1.5.2 Personal Interview

In most situations, the best method of eliciting information from individuals is by a personal interview. The interviewer personally contacts individuals selected to participate in the survey or in the experiment. Responses are recorded on *a schedule (a* questionnaire form filled out by the interviewer).

The personal interview method produces a higher response rate' than self-enumeration questionnaires and further allows the interviewer to clear up any misunderstandings about any of the questions on the schedule. But personal interviews are generally very expensive. Interviewers must be carefully selected and trained,

and sufficient remuneration must be provided to ensure that the interviewer is competent and dedicated to the chore. It is always prudent in a personal interview survey to call a selected set of respondents to ensure that they were in fact contacted (as opposed to the interviewer filling in fake responses), to ascertain if the interviewer's demeanor was appropriate, and to determine whether the interviewer may have biased responses by making gestures when stating the questions or recording the responses.

Overall, the *personal interview* method of conducting an experiment or survey, where the population units are people, is the best way to acquire data if the process is properly planned and executed, and if it can be afforded.

1.5.3 Telephone and Internet Interviews

Occasionally, it is possible to conduct an interview over the telephone or internet with the interviewer working from a schedule as in a personal interview. Polls to determine the most popular programs on television are frequently conducted in this manner. Telephone and internet interviews are usually less expensive than personal interviews, but the response rate is lower and fewer questions may be asked before the respondent tires of the proceedings. And, not everyone owns a phone or e-mail-even today.

1.6 Constructing Questionnaires and Schedules

There are three basic steps to constructing a questionnaire or schedule: (1) designing the instrument, (2) the pretest, and (3) editing the results. The construction of a questionnaire or schedule instrument is time consuming and difficult. There is a natural tendency to rush through the construction of the instrument so that the data collection process can commence. But time spent in this stage of a well-planned survey or experiment is invariably found to be extremely valuable in retrospect.

The proper construction of questionnaires is a skill which is generally developed only by experience in the use of research methodology or by on - the-job training. We will discuss only some of the basic concepts concerning the construction of a questionnaire.

1.6.1 The Design

There are basically three kinds of questions that may be asked: *dichotomous*, *multiple choice*, or *free answer*. In the dichotomous question, the respondent is asked to select one of two responses, usually "yes" and "no." For example, in a transportation study, a worker may be asked,

Did you drive a car to work this morning? YES ( ), NO ( ) .

The dichotomous question is simple and straightforward, and perhaps comes closest to decisions that respondents are used to making.

In the multiple choice question, the respondent is asked to select one of a number of responses:

What is the likelihood of your using the following services for preventive health care purposes in the next two years? *(a)* Dental check-up, (b) Eye exam, (c) General physical.

	a	*b*	*c*
Extremely unlikely	( )	( )	( )
Unlikely	( )	( )	( )
Slightly unlikely	( )	( )	( )
Not certain	( )	( )	( )
Slightly likely	( )	( )	( )
Likely	( )	( )	( )
Extremely likely	( )	( )	( )

The *multiple choice* question gives the respondent a greater range of responses to choose from, but it may also request a more qualified response than the respondent is prepared to make. For instance, a respondent may answer "yes" to the dichotomous question, "will you have a physical this year" when it is not a certain event-good intentions are not always realized. Yet, the respondent may not be able to properly conceptualize the assignment of a *likelihood* (slightly likely, likely, extremely likely) to the event, "I will have a physical this year." All too often, responses in situations like this lead to "end-loading"-selecting the response which most closely approaches a simple "yes-no" response.

In this case, the respondent would select the "extremely likely" response in place of "yes" if he or she is given the multiple choice response format, for instance.

In the free answer form, the respondent is asked to answer a question in his or her own words in essay form:

What is your opinion of the dining hall food and service?

The difficulty with the *free answer* question is in classifying the responses. This may not only be difficult and somewhat arbitrary, but it is also extremely time consuming.

In most instruments, it usually is necessary to employ all three types of questions to elicit the information required. ,

The order of the questions in the instrument can be extremely important. The questionnaire or survey should begin slowly, with easily answered questions to develop rapport with the respondent. Respondents tend to "tie" questions together and a particular ordering of questions may produce a different set of responses than another set for this reason.

The degree of directness of the questions is also important. If sensitive questions are asked directly, respondents may distort their answers. This invariably happens when a person is asked for his income. To elicit information about sensitive questions, indirect questions may be employed. For example, we may ask the respondent to indicate his salary range among a set of ranges. Later, we may ask what proportion of his monthly income is spent on food and much later, what his average monthly expenditure for food is. We may be able to determine a person's salary indirectly in this way better than by directly asking for his or her income. At the very least, we have a consistency check to determine how reliable the responses are.

It is important that the questions are stated clearly and do not bias the results. Ideally, the question should have the same meaning to every respondent in the survey or experiment. And the questions should be relatively short. Bias may arise when leading questions are used, such as:

The food in the dining hall is rotten.

Agree ( ) Uncertain ( ) Disagree ( ) .

Given to the typical college student, the response will invariably be, "agree." A less biased question might be, "The food in the dining hall is of acceptable quality."

1.6.2 The Pretest

The pre-test is an essential step in constructing a questionnaire or schedule instrument. The instrument is given to a small number of respondents to determine whether there are any problems with it. Almost always there are. There may be ambiguous questions, the ordering may require changing, and some questions may have to be asked in different forms. The time to identify difficulties with the instrument is before the full scale survey or experiment is conducted-not after.
Further, the information gathered during the pretest phase may be used to estimate statistics required for the proper planning of the statistical design of the experiment.

1.6.3 Editing

The completed questionnaires or schedules must be carefully checked and edited for errors. Often, it is possible to design in questions which represent internal consistency checks for the respondent's answers. Finding recording, transcription, or clerical errors can be very tedious work, but it is necessary if the data are going to be of value in decision making. Today, the computer is used extensively to edit data. Various computer assisted techniques have been developed to identify "outliers"-responses which are greatly different from the majority of the responses. Many outliers result from recording, transcription, or clerical errors, or from false information provided by the respondent.

1.7 Variables and Scales of Measurement

The characteristic of the population under study is called a variable if it can take on two or more different values among the population units. For instance, if we are interested in the incomes of workers in a particular industry, we may also record other characteristics about the worker as well, as for example age, race, level of education, and sex.

In this instance, the five characteristics-income, age, race, level of education, and sex-are variables in the survey or experiment.

Further, we would call income a ***dependent variable*** and the other four ***independent variables*** if we are concerned with how sex, age, level of education, and race affect income. Income is the basic variable of interest and our interest in the other variables is in their influence on income.

If we are measuring a set of variables from a population, the determination of which are dependent and independent variables is a function of the purpose of the survey or experiment. An independent variable in one study may be a dependent variable in another.

A ***quantitative variable*** is one that can be measured numerically, such as income and age. A ***qualitative variable*** is one that is nonnumeric , such as sex, race, and level of education (high school, college, graduate school, etc ...

In preparing data for analysis, we must be familiar with the four numerical scales of measurement: ***nominal, ordinal, interval*** and ***ratio***. The nominal scale applies whenever we have used numbers only to categorize outcomes of a variable. For instance, we could let a "male" be 1 and a "female" be 0, but this numerical assignment is clearly arbitrary-a female could be assigned 100 and a male, 0. The ordinal scale differs from the nominal scale in that the ordering of the numbers has meaning. An example is the responses to a multiple-response question:

Strongly agree	Agree	Uncertain	Disagree	Strongly disagree
-2	-1	0	+1	+2

The numerical assignments of -2, - 1, 0, 1, and 2 indicate the degree of agreement, but they could just as easily have been 0, 10, 100, 200, and 500, respectively. The key here is that while a -2 indicates stronger agreement than a - 1, the difference between -2 and - 1 may not be the same as between 0 and + 1. In the ***interval scale***, the relative order of the numbers is important, but so is the difference between them. This scale uses the concept of unit distance such that the difference between any two numbers may be expressed as some number of units. The interval scale requires a zero point, but its location may be arbitrary.

Good examples of interval scales are the Fahrenheit and Celsius temperature scales. Both have different zero points and unit distances. The principle of an interval scale is not violated by a change in scale or location or both. The ***ratio scale*** is used when the interval size is important and also the ratio between two numbers has meaning. By this, we mean it is appropriate to speak of one number being, say, twice as big as another. This is clearly not possible with an interval scale, where, for instance, 80°F **is** not twice as "hot" as 40°F-measured on the Celsius scale, these two temperatures are 27°C and 4°C, respectively, and 27°C is not twice 4°C. Examples of instances when ratio scales are appropriate are measurements of heights, weights, and age. Most of the statistical methods we will develop in this book require that the variable be measured at least on the interval scale.

Tutorial 1

l. Briefly describe each of the following terms:
a. A statistician.
b. Schedule.
c. Descriptive statistics.
d. Questionnaire.
e. Inferential statistics.
f. Survey.
g. Population.
h. Experiment.
i. Population characteristic.
j. Variable.
k. Census.
l. Quantitative variable.
m. Sample.
n. Qualitative variable.
q. Sampling error.
r. Dependent variable.
s. Nonsampling error.
t. Independent variable.
u. Frame.

2. Distinguish between a schedule and a questionnaire. What is each used for?

3. Distinguish between a survey and an experiment. Which is preferred and why?

4. Distinguish between primary and secondary data. Which is the most reliable? Why?

5. There are three kinds of questions that may be used in a schedule or in a questionnaire. Describe each, and discuss its advantages and disadvantages.

6. In constructing a schedule or questionnaire, there are three primary steps: design, pretest and editing. Describe each step.

7. There are four measurement scales: nominal, ordinal, interval, and ratio.

Describe each, and give an example of a survey question that may use measurements of each type.

8. For each of the following, indicate the scale of measurement: a. Red (1), Blue (0), Yellow (- 1)
b. Extremely Likely (5), Likely (4), Indifferent (3), Unlikely (2) and Extremely Unlikely (1).
c. Pressure in pounds per square inch; from 0 to ∞.
d. Volume in cubic centimeters from 0 to ∞.
e. Age in years 0 to ?
f. Salary in dollars 0 to ?
g. Rank of a state in population 1 to 50.

9. For each of the following, indicate whether it is a quantitative or qualitative variable.
a. Hair color.
b. Sales volume of an automotive firm.
c. Sex of an individual.
d. Number of persons unemployed.

10. Distinguish between sampling and nonsampling error. Which can occur in a census? Which can occur in a sample?

11. A manufacturer buys electronic parts from a supplier with the understanding that 1 percent or less of the parts are defective. In a particular shipment of 5,000 parts, the supplier finds in a sample of 100 parts that none are defective. The manufacturer decides to check the parts as well and, in another sample of 100 parts, finds that four are defective. On this basis, the manufacturer decides to reject the lot.
a. How is it possible that one sample produces 0 percent defectives while another produced 4 percent defectives?
b. Is it possible that the manufacturer is making a mistake by not accepting the shipment.

2

Frequency Distributions

2.1 Introduction

Grouping, classifying, and thus describing measurements and observations is as basic in statistics as it is in science and in many activities of everyday life. To illustrate its importance in statistics, let us consider the problem of an economist who wants to study the size of farms in the United States. Not even giving a thought to the possibility of conducting a survey of his own, since the expense would be staggering, he immediately turns to one of the many organizations that specialize in the gathering of statistical data, namely, the US. Department of Commerce. This department not only provides government agencies with statistical data needed for over-all planning and day-by-day operations, but it also makes this information available to businessmen and research workers in various fields. Like other organizations engaged in gathering statistical data, it thus faces the problem of how to present the results of its surveys in the most effective and the most usable form. With reference to the information needed by the above-mentioned economist, the Department of Commerce could print sheets containing millions of numbers, the actual sizes of all farms in the United States; it is needless to say, however, that this would not be very effective and, without some treatment, not very "usable."

When dealing with large sets of numbers, a good over-all picture and sufficient information can often be conveyed by grouping the data into a number of classes, and the Department of Commerce could, and in fact does, publish its data on the size of farms in tables like the following:

Table 2.1

Size of Farms in 1964 (acres)	Number of Farms (thousands)
Under 10	183
10-49	637
50-99	542
100-179	633
180-259	355
260-499	451
500-999	210
1,000 and over	145
Total	3156

This kind of table is called a frequency distribution (or simply a

distribution) : It shows the frequencies with which the farm sizes are distributed among the chosen classes. Tables of this sort, in which the data are grouped according to numerical size, are called numerical or ***quantitative distributions***. In contrast, tables like the one given below, in which the data are sorted according to certain categories, are called ***categorical*** or ***qualitative distributions***, as table 2.2 below:

Table 2.2

	1967 Motor Vehicle Registration (thousands)
United States	**96945**
Other North and Central America	**8900**
South America	**5490**
Europe	**65969**
Africa	**3822**
Asia	**13937**
Oceania	**5519**

Although frequency distributions present data in a relatively compact form, give a good over-all picture, and contain information which is adequate for many purposes, there are evidently some things which can be obtained from the original data that cannot be obtained from a distribution. For instance, referring to the first of the above tables, we cannot find the exact size of the smallest and largest farms, nor can we find the exact average size of the 542000 farms in the 50-99 acre group. Nevertheless, frequency distributions present raw (unprocessed) data in a more usable form, and the price which we must pay, the loss of certain information, is usually a fair exchange.

Data are sometimes grouped solely to facilitate the calculation of further statistical descriptions.

2.2 Frequency Distributions

The ***construction*** of a numerical distribution consists essentially of three steps:
(1) we must choose the classes into which the data are to be grouped,
(2) we must sort (or tally) the data into the appropriate classes, and
(3) we must count the number of items in each class.
Since the last two of these steps are purely mechanical, we shall concentrate on the first, namely, the problem of choosing suitable classifications. Note that if the data are recorded on punch-cards or tape, methods that are nowadays widely used, the sorting and counting can be done automatically in a single step.

The two things we shall have to consider in the first step are those of determining the number of classes into which the data are to be grouped and the range of values each class is to cover, that is, "from where to where" each class is to go. Both of these choices are largely arbitrary, but they depend to some extent on the nature of the data and on the ultimate purpose the distribution is to serve. The following are some ***rules*** which are generally observed:

(a) We seldom use fewer than 6 or more than 15 classes. This rule reflects sound practice based on experience; in any given example, the actual choice will have to depend on the number of observations we want to group (we would hardly group 5 observations into 12 classes), and on their range.
(b) We always choose classes which will accommodate all the data. To this end we must make sure that the smallest and largest values fall within the classification, and that none of the values can fall into possible gaps between successive classes.
(c) We always make sure that each item goes into only one class. In other words, we must avoid successive classes which overlap, that is, successive classes having one or more values in common.
(d) Whenever possible, we make the class intervals of equal length, that is, we make them cover equal ranges of values. It is generally desirable to make these ranges (intervals) multiples of 5, 10, 100, etc.., or other numbers that are easy to work with, to facilitate the tally (perhaps, mechanically) and the ultimate use of the table.

Note that the first three, but not the fourth,of these rules were observed in the construction of the farm-size distribution on page 23, assuming that the figures were rounded to the nearest acre. (Had these figures been rounded to the nearest tenth of an acre, a farm of, say, 49.6 acres could not have been accommodated, as it would have fallen between the second class and the third.) The fourth rule was violated in two ways: First, the intervals from 10 to 49 acres, 100 to 179 acres, and 260 to 499 acres, among others, cover unequal ranges of values. Second, the first and last classes are open-for all we know, the last class might include farms of a million acres or more, and if we had grouped profits and losses instead of acreages, the first class might even have included negative values. If a set of data contains a few values that are much greater (or much smaller) than the rest, open classes can help to simplify the over-all picture by reducing the number of required classes; otherwise, open classes should be avoided as they can make it impossible (or at least difficult) to give further descriptions of the data.

As we have pointed out in the preceding paragraph, the appropriateness of a classification may depend on whether the data are rounded to the nearest acre or to the nearest tenth of an acre. Similarly, it may depend on whether data are rounded to the nearest dollar or the nearest cent, whether they are given to the nearest inch, the nearest tenth of an inch, or the nearest hundredth of an inch, and so on. Thus, if we wanted to group the amounts of the sales made by a saleslady in a department store, we might use the classification given in table 2.3 below

Table 2.3

Size of Sale (dollars)
0.00 - 4.99
5.00 - 9.99
10.00 – 14.99
15.00 – 19.99
20.00 – 24.99
Etc.

And if we wanted to group the heights of children measured to the nearest tenth of an inch, we might use the classification shown in table 2.4

Table 2.4

Height (inches)
20.0 - 29.9
30.0 – 39.9
40.0 – 49.9
50.0 – 59.9
Etc.

Note that in each of these examples the nature of the data is such that a value can fall into one and only one class.

To give a concrete illustration of the construction of a frequency distribution, let us consider the following data(table 2.5) representing the scores which 150 applicants for secretarial positions in a large company obtained in an achievement test:

Table 2.5

27	**79**	**69**	**40**	**51**	**88**	**55**	**48**	**36**	**61**
53	**44**	**94**	**15**	**65**	**42**	**58**	**55**	**69**	**63**
70	**48**	**61**	**55**	**60**	**25**	**47**	**78**	**61**	**54**
57	**76**	**73**	**62**	**36**	**67**	**40**	**51**	**59**	**68**
27	**46**	**62**	**43**	**54**	**83**	**59**	**13**	**72**	**57**
82	**45**	**54**	**52**	**71**	**53**	**82**	**69**	**60**	**35**
41	**65**	**62**	**75**	**60**	**42**	**55**	**34**	**49**	**45**
49	**64**	**40**	**61**	**73**	**44**	**59**	**46**	**71**	**86**
43	**69**	**54**	**31**	**56**	**51**	**75**	**44**	**66**	**53**
80	**71**	**53**	**56**	**91**	**60**	**41**	**29**	**56**	**57**
35	**54**	**43**	**39**	**56**	**27**	**62**	**44**	**85**	**61**
59	**89**	**60**	**51**	**71**	**53**	**58**	**26**	**77**	**68**

62	**57**	**48**	**69**	**76**	**52**	**49**	**45**	**54**	**41**
33	**61**	**80**	**57**	**42**	**45**	**59**	**44**	**68**	**73**
55	**70**	**39**	**58**	**69**	**51**	**85**	**46**	**55**	**67**

Since the smallest of these scores is 17 and the largest is 94, it would seem reasonable (for most practical purposes) to choose the nine classes going from 10 to 19, from 20 to 29, . . ., and from 90 to 99. Performing the actual tally and counting the number of values falling into each class, we obtain the results shown in table 2.6 . The numbers shown in the right-hand column of this table are called class frequencies; they give the number of items falling into each class. Also, the smallest and the largest values that can go into any given class are referred to as its class limits; thus, the class limits of the above table are 10 and 19, 20 and 29, 30 and 39, and so on. More specifically, 10, 20, 30, . . ., and 90 are referred to as the lower class limits, while 19, 29, 39, ..., and 99 are referred to as the upper class limits of the respective classes.

Table 2.6

Scores	***Tally***	***Frequency***
10 -19	***1***	**1**
20 – 29	***111111***	**6**
30 – 39	***111111111***	**9**
40 – 49	***11111111111111111111111111111***	**31**
50 – 59	***1111111111111111111111111111111111111***	**42**
60 – 69	***11111111111111111111111111111***	**32**
70 – 79	***11111111111111***	**17**
80 – 89	***1111111111***	**10**
90 – 99	***11***	**2**
		150

If we are dealing with figures rounded to the nearest whole number, as in the size-of-farms distribution on page 23, the class which has the limits 10 and 49 actually contains all values between 9.5 and 49.5. Similarly, if we are dealing with measurements rounded to the nearest tenth of an inch, as in the height distribution on page 27, the class which has the limits 30.0 and 39.9 actually contains all values between 29.95 and 39.95, and the class which has the limits 40.0 and 49.9 actually contains all values between 39.95 and 49.95. It is customary to refer to these dividing lines between successive classes as the ***class boundaries***, although they are sometimes referred to

instead as the "real" ***class limits***. In order to make this concept apply also to the classes which are at the two extremes of a distribution, we simply act as if the table were continued in both directions. Thus, the first class of the above distribution of the 150 scores has the lower boundary 9.5, while the last class has the upper boundary 99.5.

It is important to remember that class boundaries should always be "impossible" values, namely, numbers which cannot occur among the values we want to group. We make sure of this by accounting for the extent to which the numbers are rounded when we choose appropriate classifications. For instance, the class boundaries of the size-of-sales distribution on page 28 are -0.005, 4.995, 9.995, 14.995, and so on. Similarly, for the distribution of the, scores, the class boundaries are 9.5, 19.5, 29.5, … , and 99.5, while the figures themselves are, of course, whole numbers. Had there been scores less than 10 in this example, we would have begun the table with the class 0-9, whose boundaries are -0.5 and 9.5.

Two other terms used in connection with frequency distributions are ***"class mark***" and "***class interval***." A *class mark* is simply the mid-point' of a class, and it is obtained by averaging the class limits (or boundaries), that is, by dividing their sum by 2. Thus, the class marks of the distribution of the scores are 14.5, 24.5, 34.5, ..., and 94.5, while those of the size-of-sales distribution, table 2.3 on page 28 are 2.495, 7.495, 12.495, and so on. A *class interval* is merely the length of a class (the range of values it can contain), and it is given by the difference between its class boundaries. If the classes of a distribution are all equal in length, their common class interval (which we refer to as the class interval of the distribution) is also given by the difference between any two successive class marks. Since 19.5 - 9.5 = 10, 29.5 - 19.5 = 10, ..., and 99.5 - 89.5 = 10, the distribution of the scores has class intervals of length 10, and we say that this is the class interval of the distribution. Note that the class interval is not given by the difference between the respective upper and lower class limits, which in our example would equal 9, and not 10.

Suppose now that in connection with the scores of the 150 applicants for secretarial positions, it is of interest to know how many fell below various levels. To provide this information, we have only to convert the distribution, table 2.6on page 30 into what is called a cumulative frequency distribution or simply a cumulative distribution. Successively adding the frequencies in the table, we thus obtain the following "less than" *cumulative distribution*, shown in table 2.7

Table 2.7

Scores		*Cumulative Frequencies*
Less than	**10**	**0**
Less than	**20**	**1**
Less than	**30**	**7**
Less than	**40**	**16**
Leas than	**50**	**47**
Less than	**60**	**89**
Less than	**70**	**121**
Less than	**80**	**138**
Less than	**90**	**148**
Less than	**100**	**150**

Note that in this table we could just as well have written "9 or less" instead of "less than 10," "19 or less" instead of "less than 20," ..., and "99 or less" instead of "less than 100."

If we successively add the frequencies starting at the other end of the distribution, we similarly get a cumulative "or more" distribution (or a cumulative "more than" distribution), which shows how many of the scores are "10 or more" (or "more than 9"), how many are "20 or more" (or "more than 19"), and so on.

Sometimes it is preferable to show what percentage of the items falls into each class, or what percentage of the items falls above or below various values. To convert a frequency distribution (or a cumulative distribution) into a corresponding percentage distribution, we have only to divide each class frequency (or each cumulative frequency) by the total number of items grouped and multiply by 100. For instance, for the size-of-farm distribution on page 25, it may be more informative to indicate that

(183/ 3,156)100 = 5 .8 per cent of the farms are under 10 acres, that

(637/3,156)100 =20.2 per cent of the farms are from 10 to 49 acres, and so on. Generally *speaking, percentage distributions are useful, especially when we want to compare two or more sets of data.*

For instance, it may well be more informative to say that the percentages of farms under 10 acres in two counties are, respectively, 5 per cent and 6 per cent, than to report that in one county 16 of 321 farms and in the other county 43 of 717 farms are under 10 acres.

So far we have discussed only numerical distributions, but the general problem of constructing categorical (or qualitative) distributions is very much the same. Again we must decide how many classes (categories) to use and what kind of items each category is to contain, making sure that all of the items are accommodated and that there are no ambiguities. Since the categories must often be selected before any data are actually obtained, sound practice is to include a category labeled "others" or "miscellaneous."

When dealing with categorical distributions we do not have to worry about such mathematical details as class limits, class boundaries, class marks, etc..; on the other hand, we now have a more serious problem with ambiguities, and we must be careful and explicit in defining what each category is to contain. For instance, if we tried to classify items sold at a supermarket into "meats," "frozen foods," "baked goods," and so oil, it would be difficult to decide where to put, for example, frozen beef pies. Similarly, if we wanted to classify occupations, it would be difficult to decide where to put a farm manager, if our table contained (without qualification) the two categories "farmers" and "managers." For this reason, it is often advisable to use standard categories developed by the Bureau of the Census and other government agencies.

2.3 Graphical Presentations

When frequency distributions are constructed primarily to condense large sets of data and display them in an "easy to digest." form, it is usually advisable to present them graphically, that is, in a form that appeals to the human power of visualization.

Some of the common graphical presentations of statistical data are :

1)Histogram, 2)Bar chart, 3)Polygon, 4)Curve, 5)Pictogram, 5)Cumulative(less) distribution and 6)Cumulative(more) distribution.

The most common among all graphical presentations of statistical data is the *histogram*, an example of which is shown in figure 2.1. A histogram is constructed by representing

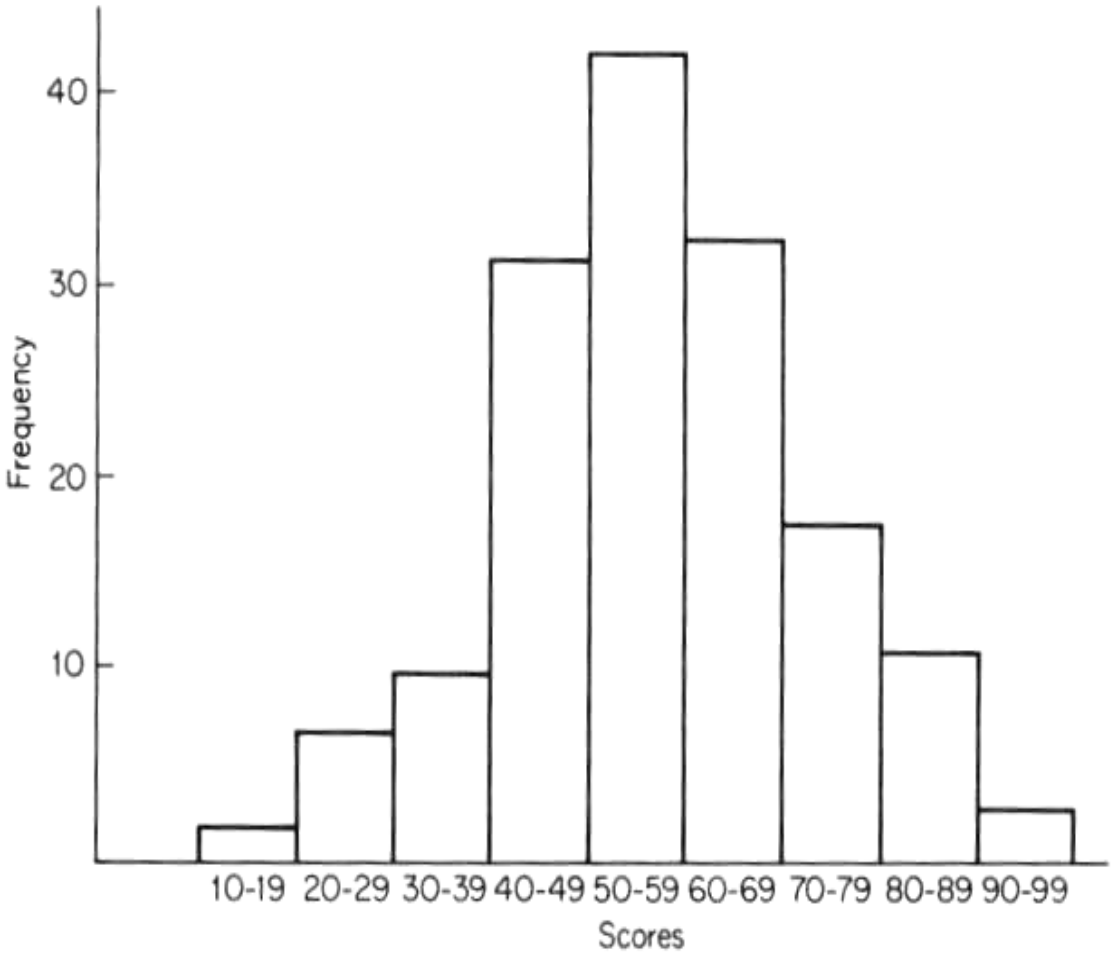


Figure 2.1. Histogram of the distribution of the scores of the 150 applicants.

measurement or observations that are grouped (in figure 2.1 the scores) on a horizontal scale, the class frequencies on a vertical scale, and drawing rectangles whose bases equal the class interval and whose heights are determined by the corresponding class frequencies. The markings on the horizontal scale can be the class limits as in figure 2.1, the class boundaries, the class marks, or arbitrary key values.

For easy readability it is generally preferable to indicate the class limits, although the bases of the rectangles actually go from one class boundary to the next. Similar to histograms are *bar charts*, like the one of figure 2.2, where the lengths of the bars are proportional to the class frequencies, but there is no pretense of having a continuous (horizontal) scale.

There are several points that must be watched in the construction of histograms. First, it must be remembered that this kind of figure cannot be used for distributions with open classes. Second, it should be noted

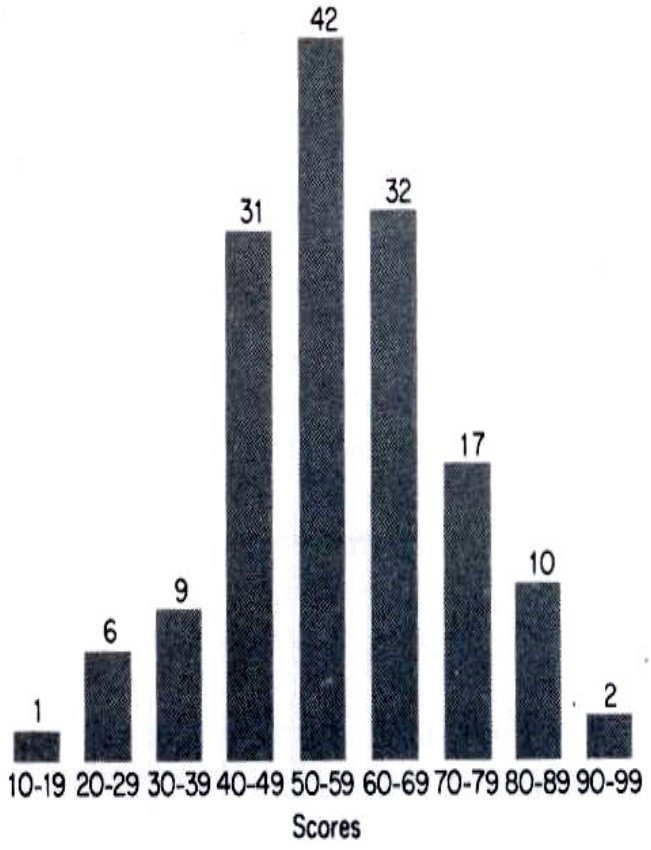


Figure 2.2. Bar chart of the distribution of the scores of the 150 applicants.

that the picture presented by a histogram can be very misleading if a distribution hits unequal classes and no suitable adjustments are made. To illustrate this point, let us regroup the distribution of the 150 scores by combining all those from 60 to 79 into one class. Thus, the new distribution is given by the following table

Table 2.8

Scores	*Frequency*
10-19	**1**
20-29	**6**
30-39	**9**
40-49	**31**
50-59	**42**
60-79	**49**
80-89	**10**
90-99	**2**

and its *histogram* (with the class frequencies represented by the heights of the rectangles) is shown in figure 2.3. This figure gives the impression that

just about half the scores fall on the interval from 60 to 79, where as

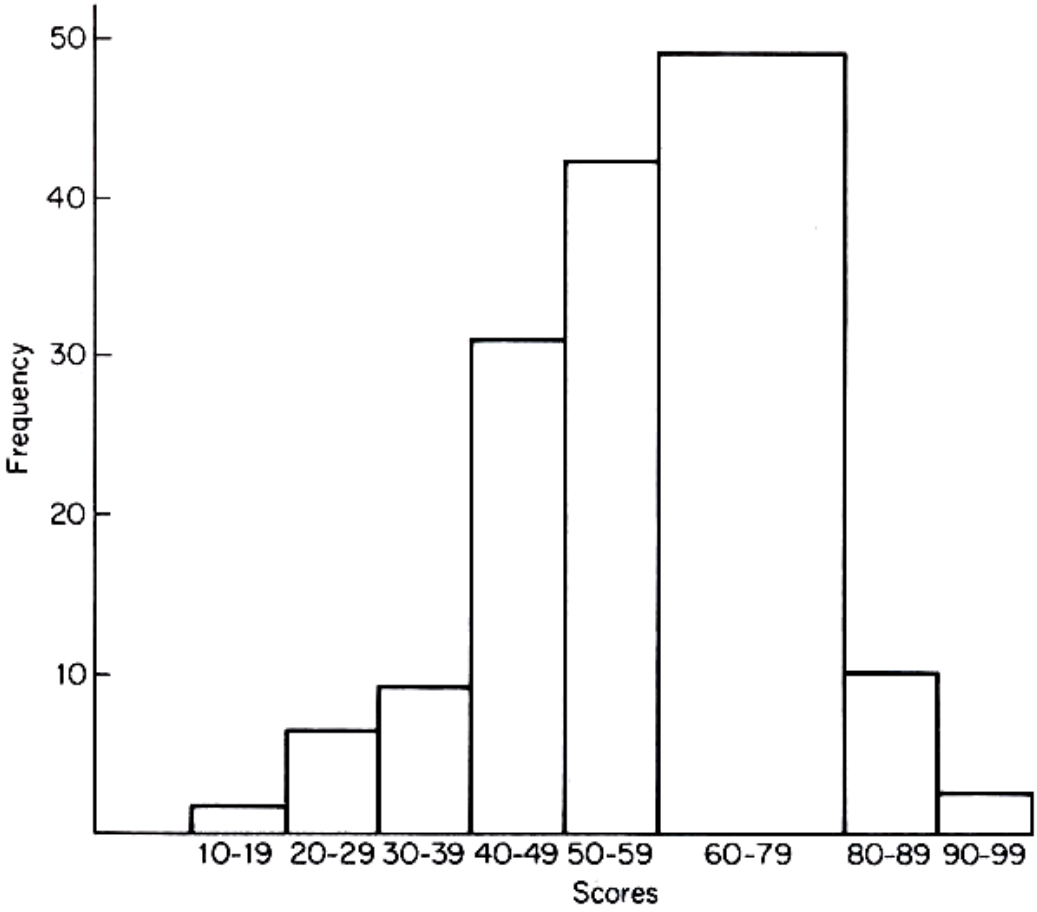


Figure 2.3. Incorrectly modified histogram of the distribution of the scores.

the correct proportion is close to 1/3,49/150 to be exact. This error is due to the fact that when we compare the size of rectangles, triangles, and other plane figures, we instinctively compare their areas and not their sides. In order to correct for this, we simply draw the rectangles of the histogram so that the class frequencies are represented by their areas, and not by their heights. In figure 2.4 we accomplished this by reducing the height of the rectangle representing the class 60-79 to half of what it was in figure 2.3.

The practice of representing class frequencies by means of areas is especially important if histograms are to be approximated with smooth curves. For instance, if we wanted to approximate the histogram of figure 2.1 with a smooth curve, we could say that the number of scores exceeding 69 is given by the shaded area of figure 2.5. Clearly, this area is approximately equal to the sum of the areas of the corresponding three rectangles.

An alternate, though less widely used, form of graphical presentation is the *frequency polygon* (see figure 2.6). Here the class frequencies are plotted at the class marks and the successive points are connected by means of straight lines. Note that we added classes with zero frequencies

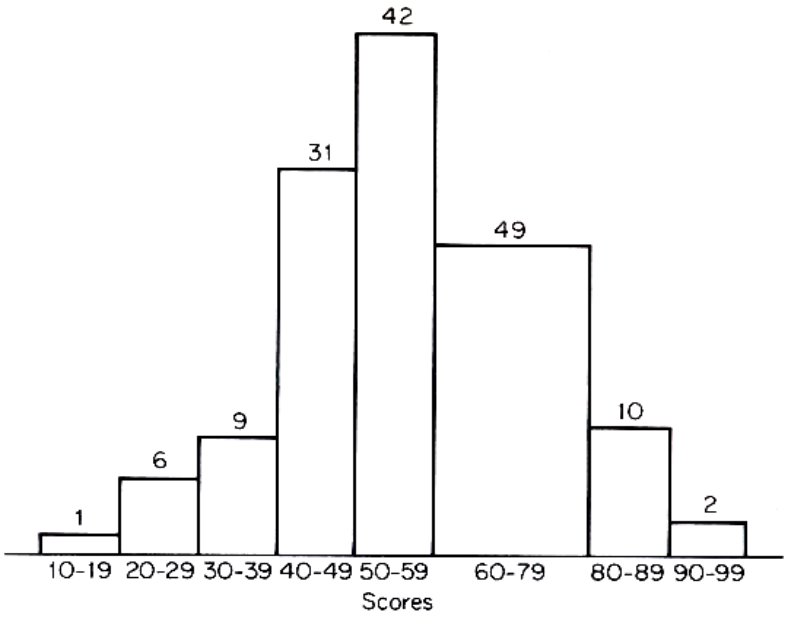


Figure 2.4. Correctly modified histogram of the distribution of the scores.

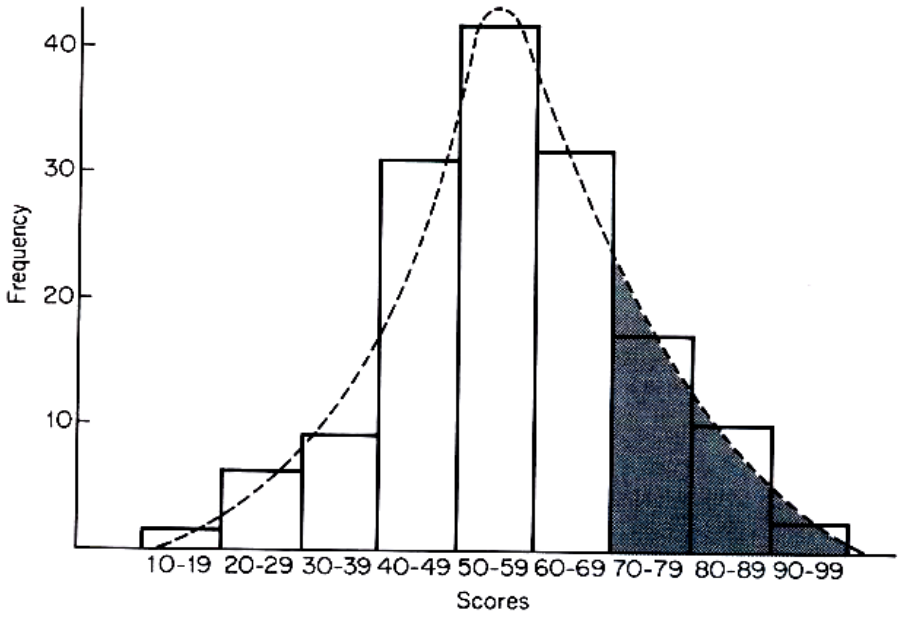


Figure 2.5. Histogram of the distribution of the scores approximated with a smooth curve.

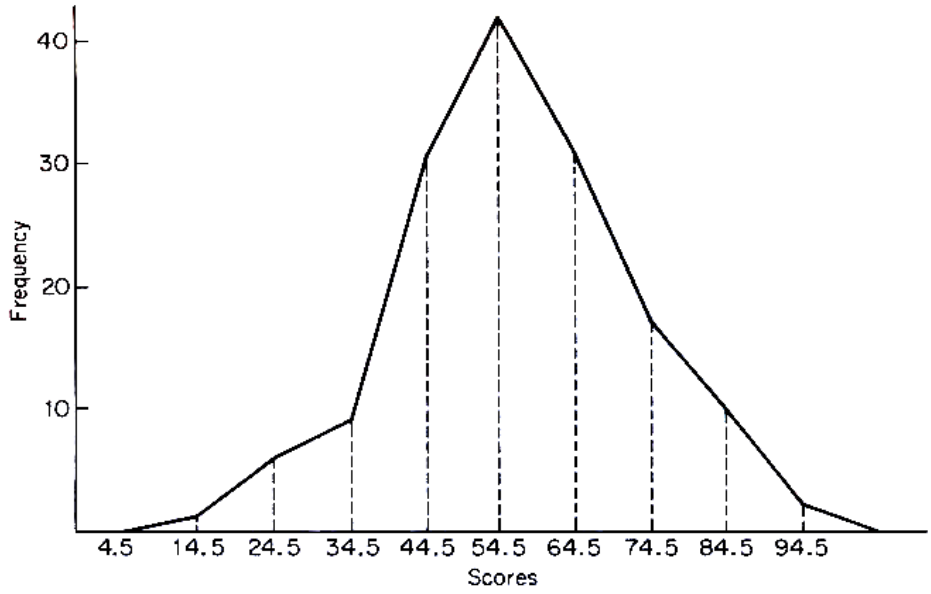


Figure 2.6. Frequency polygon of the distribution of the 150 scores.

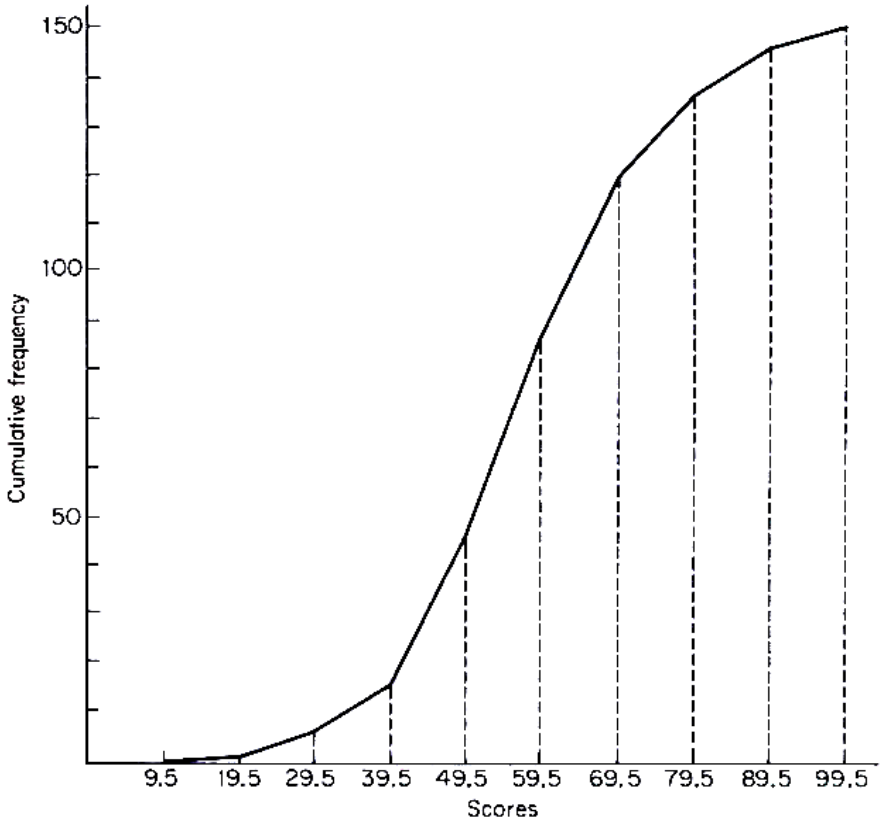


Figure 2.7. Ogive of the distribution of the 150 scores.

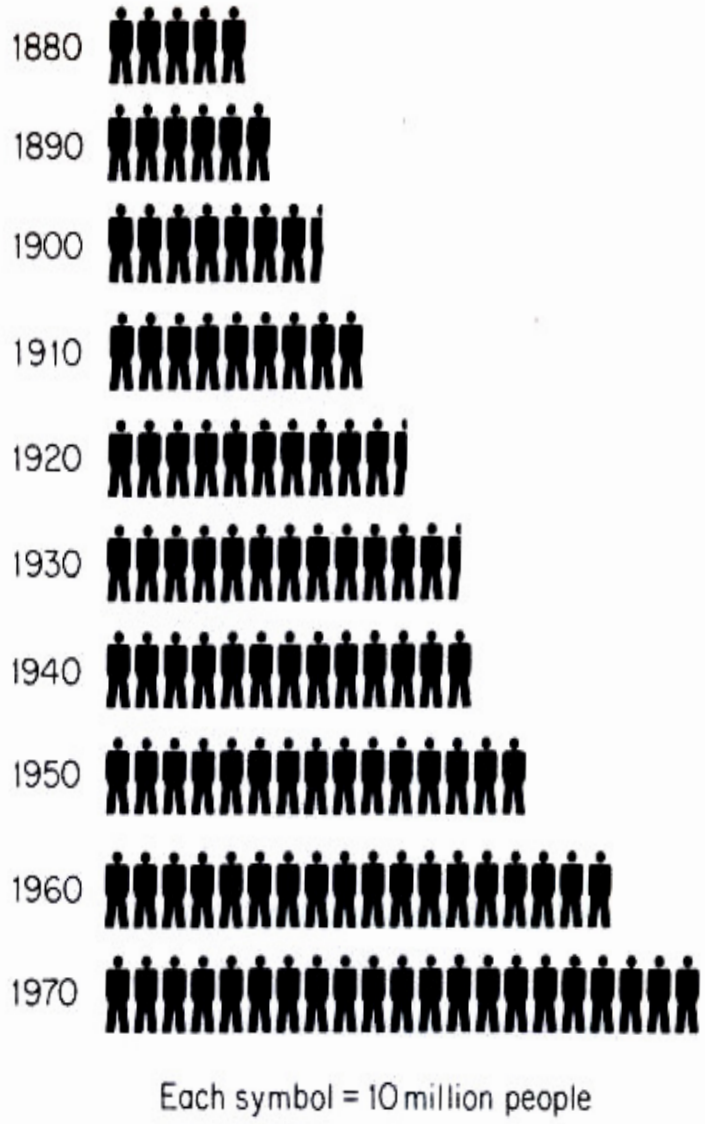


Figure 2.8. Pictogram of the population of the United States.

at both ends of the distribution in order to "tie down" the graph to the horizontal scale.

If we apply the same technique to a *cumulative distribution*, we obtain what is called an *ogive*. Note, however, that now the cumulative frequencies

are not plotted at the class marks-it stands to reason that the cumulative frequency corresponding, say, to "less than 20" in our example should be plotted at 20, or preferably at the class boundary of 19.5, since "less than 20" actually includes everything up to 19.5. figure 2.7 shows an ogive representing the cumulative "less than" distribution of the scores of the 150 applicants.

Although the visual appeal of histograms, frequency polygons, and ogives exceeds that of frequency tables, there are ways in which distributions can be presented even more dramatically and probably also more effectively. We are referring here to the various kinds of *pictograms (pictorial presentations*)(see, for example, figure 2.8) with which the reader must surely be familiar through newspapers, magazines, advertising, and other sources. The number of ways in which distributions (and other statistical data) can be displayed pictorially is almost unlimited, depending only on the imagination and artistic talent of the individual preparing the presentations.

Tutorial 2

1. Decide for each of the following quantities whether it can be determined on the basis of the distribution of the 150 scores on page 30; if possible, give a numerical answer:
   (a) The number of scores which were at least 50.
   (b) The number of scores which were greater than 50.
   (c) The number of scores which were 80 or less.
   (d) The number of scores which were less than 80.
   (e) The number of scores which were more than 90.
   (f) The number of scores which were greater than 39 but at most 69.

2. If the amounts paid for the repairs of cars damaged in accidents are grouped into a frequency table with the classes \$0.00-\$99.99, \$100.00\$199.99, \$200.00-\$299.99, \$300.00-\$399.99, \$400.00-\$499.99, and \$500.00 or more, decide for each of the following quantities whether it can be determined on the basis of this distribution:
   (a) How many of the amounts were less than \$200.00.
   (b) How many of the amounts were at least \$200.00.
   (c) How many of the amounts were more than \$200.00.
   (d) How many of the amounts were \$200.00 or more.

3. The following is the distribution of the weekly earnings of 1, 216 secretaries in the Phoenix, Arizona, metropolitan area in March, 1969:

Weekly Earnings	***Number of Secretaries***
Under \$80	**21**
\$80- \$99	**296**
\$100-\$119	**494**
\$120-\$139	**247**
\$140-\$159	**119**
\$160 and over	**39**

Decide for each of the following quantities whether it can be determined on the basis of this distribution; if possible give a numerical answer:

(a) The number of secretaries with weekly earnings of at least \$120.
(b) The number of secretaries with weekly earnings of more than \$120.
(c) The number of secretaries with weekly earnings of more than \$180.
(d) The number of secretaries with weekly earnings of less than \$100.
(e) The number of secretaries with weekly earnings of at most \$100.
(f) The number of secretaries with weekly earnings of at least \$60.

4. The number of students absent from school each day are grouped into a distribution having the classes 3-10, 11-18, 19-26, 27-34, and 35-42. Find (a) the limits of each class, (b) the class boundaries, and (c) the class marks.

5. The following is the distribution of the actual weight (in ounces) of 50 "one-pound" bags of coffee, which a grocery clerk filled from bulk stock:

Weight	***Number of bags***
15.5 – 15.6	3
15.7 – 15.8	9
15.9 – 16.0	17
16.1 – 16.2	14
16.3 – 16.4	6
16.5 – 16.6	1

Find (a) the limits of each class, (b) the class marks, and (c) the class boundaries.

6. The weights of certain laboratory animals, given to the nearest tenth of an ounce, are grouped into a table having the class boundaries 11.45, 13.45, 15.45, 17.45, and 19.45 ounces. What are the limits of the four classes of this distribution?

7. The class marks of a distribution of temperature readings, given to the nearest degree Fahrenheit, are 113, 128, 143, 158, and 173. Find the class boundaries of this distribution, and also the class limits.

8. Class limits and class boundaries have to be interpreted very carefully when we are dealing with ages, for the age group from 5 through 9, for example, includes all those who have passed their fifth birthday but not

yet reached their tenth. Taking this into account, what are the boundaries and the class marks of the following age groups: 10-19, 20-29, 30-39, and 40-49.

9. A study of air pollution in a city yielded the following daily readings of the concentration of sulfur dioxide (in parts per million):

.04 .11 .05 .01 .15 .12 .19 .06 .13 .03
.18 .01 .08 .11 .08 .14 .02 .14 .08 .10
.17 .09 .14 .07 .13 .11 .09 .05 .15 .08
.06 .05 .12 .10 .27 .12 .16 .10 .09 . I 5
.07 .10 .17 .13 .20 .18 .11 .17 .14 .04
.22 .11 .09 .02 .12 .16 .15 .12 .13 .07
.05 .14 .04 .16 .19 .10 .06 .03 .16 .13
.18 .13 .11 .09 .06 .23 .11 .12 .07 .11

(a) Group these data into a table having the classes .00-.04, .05--.09, .10-.14, .15-.19, .20-.24, and .25-.29.

(b) Convert the distribution obtained in (a) into a cumulative "less than" distribution.

(c) Construct a histogram of the distribution obtained in (a).

(d) Draw an ogive of the cumulative "less than" distribution obtained in (b) and use it to read off (roughly) the value below which we should find the lowest, half of the data

10. The following are the number of customers a restaurant served for lunch on 120 week days

50	**64**	**55**	**51**	**60**	**41**	**71**	**53**	**63**	**64**	**46**	**59**
66	**45**	**61**	**57**	**65**	**62**	**58**	**65**	**55**	**61**	**50**	**55**
53	**57**	**58**	**66**	**53**	**56**	**64**	**46**	**59**	**49**	**64**	**60**
58	**64**	**42**	**47**	**59**	**62**	**56**	**63**	**61**	**68**	**57**	**51**
61	**51**	**60**	**59**	**67**	**52**	**52**	**58**	**64**	**43**	**60**	**62**
48	**62**	**56**	**63**	**55**	**73**	**60**	**69**	**53**	**66**	**54**	**52**
56	**59**	**65**	**60**	**61**	**59**	**63**	**56**	**62**	**56**	**62**	**57**
57	**52**	**63**	**48**	**58**	**64**	**59**	**43**	**67**	**52**	**58**	**47**
63	**53**	**54**	**67**	**57**	**61**	**65**	**78**	**60**	**66**	**63**	**58**
60	**55**	**61**	**59**	**74**	**62**	**49**	**63**	**65**	**55**	**61**	**54**

(a) Group these data into a table having the classes 40-44, 45-49, 50-54, 55-59, 60-64, 65-69 ,70-74 and 75- 79.

(b) Convert the distribution obtained in (a) into a cumulative "less than" distribution.

(c) Construct a histogram of the distribution obtained in (a).

(d) Draw an ogive of the cumulative "less than" distribution obtained in (b) and use it to read off (roughly) the value below which we should find the lowest, half of the data .

3

Measures of Location

3.1 Introduction

Descriptions of statistical data can be quite brief or quite elaborate, depending partly on the nature of the data themselves, and partly on the purpose for which they are to be used. Sometimes, we even describe the same set of data in several different ways. To draw an analogy, a large motel might describe itself to the public as having luxurious facilities, a heated swimming pool, and TV in every room; on the other hand, it might describe itself to the fire department by giving the floor space of each unit, the number of sprinklers, and the number of employees. Both of these descriptions may serve the purpose for which they are designed, but they would hardly satisfy the State Corporation Commission in passing on the owner's application for issuing stock. This would require detailed information on the management of the motel, various kinds of financial statements, and so on.

Whether we describe things statistically or whether we simply describe them verbally, it is always desirable to say neither too little nor too much. Thus, it may sometimes be satisfactory to present data simply as they are and let them "speak for themselves"; in other instances it may be satisfactory to group, classify, and present them using the methods of Chapter 2. However, most of the time it is necessary to summarize them further by means of one or more well-chosen descriptions. In this chapter and in chapter 4 we shall concentrate mainly on two kinds of descriptions ,called measures of location, and measures of variation.

The *measures of location* we shall study in this chapter are also referred to at times as "*measures of central tendencies*," "measures of central values," and "measures of position." Except for some of the measures discussed in Section 3.4, they may also be referred to crudely as "averages" in the sense that they provide numbers that are indicative of the "center," "middle," or the "most typical" of a set of data.

When we said that the choice of a statistical description depends partly on the nature of the data themselves, we were referring among other things to the following distinction: if a set of data consists of all conceivably possible (or hypothetically possible) observations of a certain phenomenon, We refer to it as a population; if it contains only part of these observations, we refer to it as a sample. The qualification "hypothetically possible" was added to take care of such clearly hypothetical situations where, say, twelve flips of a coin are looked upon as a sample from the population of all

possible flips of the coin, or where we shall want to look upon the weights of eight 30-day-old calves as a sample of the weights of all (past, present, and future) 30-day-old calves. In fact, we often look upon the results obtained in an experiment as a sample of what we might obtain if the experiment were repeated over and over again.

In this chapter and the next we shall limit ourselves to methods of description without making generalizations, but it is important even here to distinguish between samples and populations. As we have said before, the kind of description we may want to use will depend on what we intend to do later on, whether we merely want to present facts about populations or whether we want to generalize from samples. We shall, thus, begin in this chapter with the practice of using different symbols depending on whether we are describing samples or populations; in Chapter 4 we shall carry this distinction one step further by even using different formulas.

3.2 The Mean

There are many problems in which we have to represent data by means of a single number which, in its way, is descriptive of the entire set. The most popular measure used for this purpose is what the layman calls an "average" and what, in statistics, is called an arithmetic mean., or simply a mean. We gave the word "average," in quotes because it generally has a loose connotation and different meanings-for example, when we speak of a batting average, an average housewife, a person with average taste, and so on.

Definition 3.1
Arithmetic mean

The arithmetic mean of a set of n numbers is defined simply as their sum divided by n.

Example 3.1

Given that the total attendance at major league baseball games in the years 1965, 1966, 1967, and 1968 was, respectively, 22.4, 25.2, 23.8, and 23.0 million, we find that the mean, namely, the "average" annual attendance for these foul: years was (22.4+25.2+23.8+23.0)/4= 23.6 million

In order to develop a simple formula for the mean that is applicable to any set of data, it will be necessary to represent the figures (measurements or observations) to which the formula is to be applied with some general

symbols such as x, y, or z. In the above example, we could have represented the annual attendance figures with the letter x and referred to the four values as x_1 (x sub-one), x_2 (x sub-two), x_3, and x_4. More generally, if we have n measurements which we designate x_1, x_2, x_3, ..., and x_n,, we can write

$$\text{Mean } \bar{x} = (x_1 + x_2 + \ldots\ldots + x_n)/n$$

This formula is perfectly general and it will take care of any set of data, but it is still somewhat cumbersome. To make it more compact, we introduce the symbol $\sum$ (capital sigma, the Greek letter for S), which is simply a mathematical shorthand notation indicating the process of summation or addition. If we write $\sum x$, this represents the "sum of the x's ," and we have

$$\bar{x} = \sum x_i/n$$

Using the sigma notation in this form, the number of terms to be added is not stated explicitly; it is tacitly understood, however, to refer to all the X's with which we happen to be concerned. . For a further discussion of the use of subscripts and the $\sum$ notation, we shall finish simplifying our notation by assigning a special symbol to the mean itself. If we look upon the x's as a sample, we write their mean as $\bar{x}$(x-bar); if we look upon them as a population, we write their mean as μ .If we refer to sample data as y's or z's, we correspondingly write their means as y or z. To further emphasize the distinction between samples and populations, we denote the number of values in a sample, the sample size, with the letter n and the number of values in a population, the population size, with the letter N. We thus have the formulas

$$\bar{x} = \sum x_i/n \quad , \quad \mu = \sum x_i/N \quad ,$$

depending on whether we are dealing with a sample or a population. In order to distinguish between descriptions of samples and descriptions of populations, statisticians not only use different symbols, but they refer to the first as statistics and the second as parameters. Hence, we say that $\bar{x}$ is a statistic and that μ is a parameter.

The popularity of the mean as a measure describing the "middle" or "center" of a set of data is not just accidental. Anytime we use a single number to describe a set of data, there are certain desirable properties we must keep in mind.

Thus, some of the noteworthy ***properties*** of the ***mean*** are:
(1) it is familiar to most persons, although they may not call it by this name,
(2) it always exists, that is, it can be calculated for any kind of numerical data,
(3) it is always unique, or in other words, a set of data has one and only one mean,
(4) it takes into account each individual item ,
(5) it lends itself to further statistical manipulation , (it is possible to combine the means of several sets of data into an over-all mean without having to refer back to the original raw data), and
(6) it is relatively reliable in the sense that it does not vary too much when repeated samples are taken from one and the same population, at least not as much as some other kinds of statistical descriptions.

This question of reliability is of fundamental importance when it comes to problems of estimation, hypothesis testing, and making predictions, and we shall have a good deal more to say about it later in this book.

Since the computation of means is quite easy, involving only addition and one division, there is usually no need to look for short-cuts or simplifications. However, if the numbers are unwieldy, that is, if each number has many digits, or if the sample (or population) size is very large, it may be advantageous to group the data first and then compute the mean from the resulting distribution. Another reason why we shall investigate the problem of obtaining means from grouped data is that published data are very often available only in the form of distributions.

3.3 The mean of a distribution

To obtain a formula for the mean of a distribution, let us write the successive class marks as x_1,, x_2, ..., x_k (assuming that there are k classed and the corresponding class frequencies as f_1, f_2, . . ., f_k . The total that goes into the numerator of the formula for the mean is thus obtained by adding f_1 times the value x_1, f_1 times the value x_2, . . ., and f_k. times the value x_k; in other words, it is equal to $x_1f_1 + x_2f_2 +. . . + x_kf_k$. Using the $\sum$ notation, we can now write the formula for the mean of a distribution .

Definition 3.2
Mean of a distribution

$$\bar{X}=\sum x_i \cdot f_i / \sum f_i,$$

where n equals $f_1 + f_2 +. ..+ f_k$, the sum of the class frequencies, or $\sum$ fi. (When dealing with a population instead of a sample, we have only to substitute μ for x in this formula and N for n.)

Example 3.2

To illustrate the calculation of the mean of a distribution, let us refer again to the distribution of the scores of the 150 applicants on chapter 2. Writing the class marks in the second column, we get

	Class Marks X	Frequencies f	Products x. f
10-19	**14.5**	**1**	**14.5**
20-29	**24.5**	**6**	**147.0**
30-39	**34.5**	**9**	**310.5**
40-49	**44.5**	**31**	**1379.5**
50-59	**54.5**	**42**	**2289.0**
60-69	**64.5**	**32**	**2064.0**
70-79	**74.5**	**17**	**1266.5**
80-89	**84.5**	**10**	**845.0**
90-99	**94.5**	**2**	**189.0**
	Total	**150**	**8505.0**

and it follows that the mean of the distribution is

$\bar{x}$=8505.0/150=56.7

It is of interest to note that the mean of the original raw is 8500/150=56.67 so that the difference between the two means is extremely small.

3.4 The coding method

The calculation of the mean of the distribution of the 150 scores was fairly easy because the frequencies were all small. Even so, the calculations can be simplified by performing a change of scale; that is, we replace the class marks with numbers that are easier to handle. This is also referred to as "coding," and in our example, we might replace the class marks of the distribution of the scores with the consecutive integers -4, -3, -2, -1, 0, 1, 2, 3, and 4. Of course, when we do something like this, we also have to account for it in the formula we use to calculate the mean. Referring to the new (coded) class marks as u's, it can easily be shown that the formula for the mean of a distribution becomes

Definition 3.3

Coding (shortcut) Mean

The coding (shortcut) mean is given by

$$\bar{X} = x_0+(\sum u_i \cdot f_i/n) \cdot c$$

where xo is the class mark (in the original scale) to which we assign 0 in the new scale, c is the class interval, n is the number of items grouped, and $\sum$ ui*fi is the sum of the products obtained by multiplying each of the coded class marks by the corresponding frequency.

Example 3.3

Illustrating this short-cut technique by recalculating the mean of the distribution of the scores of the 150 applicants, we obtain

Class Marks	ui	fi	ui.f
14.5	-4	1	-4
24.5	-3	6	-18
34.5	-2	9	-18
44.5	-1	31	-31
54.5	0	42	0
64.5	1	32	32
74.5	2	17	34
84.5	3	10	30
94.5	4	2	8
Total		150	33

$\overline{X} = 54.5+(33/150)10=56.7,$

should be noted that this agrees with the result obtained earlier; the short-cut formula does not entail any further approximation, and it should always yield the same result as the formula of definition 3.2.

Unless one can use an automatic computer, the short-cut method will generally save a good deal of time; about the only time that the short-cut method will not provide appreciable savings in time and energy is when the original class marks are already easy-to-use numbers. In order to reduce the work to a minimum, it is generally advisable to put the zero of the u-scale near the middle of the distribution, preferably at a class mark having one of the highest frequencies.

Remark 3.1:

A fact worth noting is that this short-cut method cannot be used for distributions with unequal classes, although there exists a modification which makes it applicable also in that case. Neither the short-cut formula nor the formula on definition 3.2 is applicable to distributions with. open classes; the means of such distributions cannot be found without going back to the raw data or making special assumptions about the values which fall into an open class.

Tutorial 3.1

1. Suppose we are given the high temperature recorded each day of the year 1972 in Atlanta, Georgia. Give one illustration each of a situation where these data would be looked upon (a) as a population, and (b) as a sample.
2. Suppose that the final election returns from a given county show that the two candidates for a certain office received, respectively, 16,283 and 13,559 votes. What office might these candidates be running for so that we can look upon these figures (a) as a sample and (b) as a population?
3. The dean of a college has in his files a complete record of how many A's, B's, C's, etc.., each instructor gave to his students during the academic year 1971-72. Give one illustration each of a problem (situation) in which the dean would look upon this information (a) as a sample and (b) as a population.
4. The following are the speeds (in miles per hour) at which 25 cars were timed on the San Bernardino Freeway in early-morning traffic: 52, 56, 54, 78, 71, 66, 69, 60, 70, 53, 55, 62, 67, 60, 56, 72, 73, 61, 68, 59, 67, 66, 67, 73, and 65. Find the mean of these speeds and comment on the (misleading?) argument that "on the average cars do not exceed the speed limit of 65 miles per hour on this freeway in early-morning traffic."
5. The following are the monthly water bills which a resident of Scottsdale, Arizona, received in 1971: $10.26, $9.29, $11.24, $12.22, $19.07, $21.03, $22.50, $26.41, $18.09, $23.96, $16.18, and $15.60. Find the mean, namely, the average water bill this person paid per month in 1971.
6. The following are the number of seconds which 16 insects survived after being sprayed with a certain insecticide: 121, 115, 79, 52, 102, 126, 81, 65, 109, 119, 115, 121, 103, 75, 59, and 110.
   (a) Calculate the mean of these 16 measurements.
   (b) Recalculate the mean of these 16 measurements by first subtracting 100 from each value, finding the mean of the numbers thus obtained, and then adding 100 to the result.

(What general simplification does this suggest for the calculation of means?)

7. Twenty-four cans of a floor wax, randomly selected from a large production lot, have the following net weights (in ounces): 12.0, 11.9, 12.2, 12.0, 11.9, 12.0, 12.0, 12.1, 11.8, 12.0, 12.0, 12.1, 11.9, 11.9, 12.2, 12.1, 12.0, 11.9, 11.9, 12.1, 12.0, 12.0,11.9, and 12.0.
   (a) Calculate the mean of these 24 weights.
   (b) Recalculate the mean of these 24 weights by first subtracting 12.0 from each value, finding the mean of the numbers thus obtained, and then adding 12.0 to the result. (What general simplification does this suggest for the calculation of means?)

8. The following are the number of twists that were required to break 20 forged alloy bars: 37, 29, 34, 21, 54, 38, 30, 26, 48, 37, 24, 33, 39, 51, 44, 38, 35, 29, 46, and 31. Find the mean of these values.

9. In business and economics, there are many problems in which we are interested in index numbers, that is, in measures of the changes that have taken place in the prices (quantities, or values) of various commodities. In general, the year or period we want to compare by means of an index number is called the given year or given period, while the year or period relative to which the comparison is made is called the base year or base period. Furthermore, given-year prices are denoted p_n base-year prices are denoted po, and the ratio p_n / p_o for a given commodity is called the corresponding price relative. A very simple kind of index number is given by the mean of the price relatives of the commodities with which we are concerned, multiplied by 100 to express the index as a percentage.
(a) Find the mean of the price relatives comparing the 1969 prices of the given processed fruits and vegetables (in cents) with those of 1965:

	1965	1969
Fruit cocktail, No. 303 can	**26.1**	**27.9**
Pears , No.21 can	**47.0**	**50.9**
Frozen orange juice, 6 oz,	**23.7**	**24.3**
Pears , No.303 can	**23.7**	**24.6**
Tomatoes, No. 303 can	**16.1**	**19.6**
Frozen broccoli, 10 oz,	**26.4**	**27.6**

(b) Find the mean of the price relatives comparing the following 1967 prices with those of 1960, where all prices are in cents per pound:

	1960	1967
Copper	**32.4**	**38.6**
Lead	**11.9**	**14.0**
Zinc	**12.9**	**13.8**

10. If we substitute q's for p's in the index number of Exercise 9. where given-year quantities (produced, sold, or consumed) are denoted q_n and base-year quantities are denoted q_0 , we obtain a corresponding quantity index. Given the following data in thousands of short tons, find the mean of the quantity relatives comparing the 1967 production figures with those of 1960:

	1960	1967
Copper	**1080**	**954**
Lead	**247**	**317**
Zinc	**435**	**549**

11. Another way of obtaining an index comparing given-year prices with a corresponding set of base-year prices (see Exercise 9) is to average the two sets of prices separately, take the ratio of the two means, and then multiply by 100 to express the index as a percentage. Canceling denominators, the formula for such a simple aggregative index is thus

3.5 The Mode

Definition 3.4

Mode

The ***mode***, denoted by mo, of a set n observations $x_1, x_2, \ldots, x_n$ (or of a frequency table)is the value of X which occurs with greatest frequency.

Example 3.4

Find the mode for the following observations: 3, 7, 3, 5, 2, 8

Solution:
mo=3

Example 3.5

Compute the mode for the following 12 numbers:

2 , 3, 2, 5, -1, -2, -1, 2, -1, 5, 0, 8

Solution:
mo_1=-1 , mo_2=2
In this case, we will say that the set of numbers is bimodal.

Example 3.6

Determine the mode of the six measurements: 2, 3, -1, 4, 0, 1

Solutions:
In the case, we will say that the set of numbers does not have a mode.

3.6 The Median

To avoid the difficulty met on section 3.2, where we showed that an extreme value (perhaps, a gross error) can have a pronounced effect on the mean, we sometimes describe the "middle" or "center" of a set of data with other kinds of statistical descriptions.

One of these is the median, which is defined simply as:

Definition 3.5

Median

The *median* of a set of data ,is the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an increasing or decreasing order of magnitude.

If we have an odd number of items, there is always a middle item whose value is the median. For example, the median of the five numbers 5, 10, 2, 7, and 8 is 7, as can easily be verified by first arranging these numbers according to size, and the median of the nine numbers 3, 5, 6, 9, 9, 10, 10, 12, and 13 is 9. Note that there are two 9's in this last example and that we do not refer to either of them as the median. The median is a number and not an item, namely, the value of the middle item. Generally speaking, if there are n items and n is odd, the median is the value of the (n + 1)/2 th. largest item. Thus, the median of 25 numbers is given by the value of the (25+1)/2=13th largest, the median of 49 numbers is given by the value (49+1)/2=25 th largest, and the median of 81 numbers is given by the value of the (81+1)/2=41 st largest.

If we have an even number of items, there is never a middle item, and the median is defined as the mean of the values of the two middle items. For instance, the median of the six numbers 3, 6, 8, 10, 13, and 15 (which are already ordered according to size) is (8+10)/ 2 = 9. It is halfway between the two middle values (here the 3rd and the 4th) and, if we interpret it correctly, the formula (n+1)/2 again gives the position of the median.

For the six given numbers the median is, thus, the value of the (6+1)/2=3.5th largest, and we interpret this as "halfway between the values of the third and the fourth." Similarly, the median of 100 numbers is given by the value of the (100+1)/2= 50.5 th largest item, or halfway between the values of the 50th and tile 51st.

It is important to remember that the formula (n+1)/2 is not a formula for the median, itself; it merely tells us the position of the median, namely, the number of items we have to count until we reach the item whose value is the median (or the two items whose values have to be averaged to obtain the median).

To find the median of a distribution with a total frequency of n, we must, so to speak, count n/2 items starting at either end and use def.3.6.

Definition 3.6
Median

The *median* of grouped data denoted by m is defined by
m=(Lower boundary of the class containing the median)+(((n/2)-cum.. frequency before the median class)/ frequency of median class) **.** c

Example 3.7
To illustrate this procedure , let us refer again to the distribution of the 150 scores since n=150 in this example , we will have to count n/2= 75 items from either end. Beginning at the bottom of the distribution, we find that 47 of the values are less than 50 while 89 are less than 60, so that the median must fall into the class whose limits are 50-59. Since 47 of the values fall below this class, we must count another 75 - 47 = 28 of its 42 values, and we accomplish this by adding 42 of the class interval of 10 to 49.5, the lower boundary of the class. (We add 42 of the class interval because we want to count 28 of the 42 values contained in this class.) We thus get

$$m= 49.5 +(28/42) \cdot 10=56.2$$

rounded to one decimal.

Generally speaking, if L is the lower boundary of the class containing the median, f is its frequency, c the class interval, and j the number of items we still lack when reaching *L,* then the ***median of the distribution*** *is* given by the formula

$$m=L+(j/f) \cdot c$$

It is possible, of course, to arrive at the median of a distribution by starting at the other end and *subtracting* an appropriate fraction of the class interval from the upper boundary U of the class into which the median must fall. For the distribution of the scores we thus obtain

$$m=59.5+(14/42) \cdot 10=56.2$$

and the two answers are identical, as they should be.

3.7 Other Numerical Measures

There are numerous representative measures other than the mean, median, and mode. The geometric mean is commonly used in business problems to describe the "average" of ratios. Although the geometric mean is not as important as the three principal representative measures (mean, median, and mode).

3.7.1 Geometric mean

Definition 3.7

Geometric mean

The *geometric* mean of a set of n measurements, $x_1, x_2,...,x_n$ denoted by GM is defined by

$$GM = [x_1 \cdot x_2 \cdot x_3 \ldots x_n]^{1/n}$$

The geometric mean is the nth root of the product of all the measurements. It is not as easily computed as the arithmetic mean-the computation is eased somewhat by taking logarithms of both sides of above equation

$$\log_{10}(GM) = 1/n \ [\log x_1 + \log x_2 + \ldots + \log x_n]$$

it is apparent that the geometric mean can be computed by taking the antilog of the arithmetic mean of the logs of the measurements.

Example 3.8

Determine the geometric mean of the three measurements:

$x_1 = 2$, $x_2 = 4$ and $x_3 = 8$

Solution:

$$GM = ((2)(4)(8))^{1/3} = 4.$$

Example 3.9

Find the arithmetic and geometric mean of 100, 100, 100 and 1000.

Solution:

$$\overline{X} = (100+100+100+1000)/4 = 325$$

Geometric mean: log (GM) = 1/4[1og100 + 1og100 + 1og100 + Iog1000]

$$= 1/4[2+2+2+3] = 1/(4*9) = 2.25$$

$$GM = (10)^{2.25} = 177.8.$$

The above example illustrates the fact that the geometric mean is less affected by one (or two) extremely large (or small) values than is the arithmetic mean. Unfortunately, the geometric mean is neither easy to compute nor amenable to use for statistical inferences. It is very useful, however, in averaging ratios-a process that frequently arises in computing cost-of-living or other index numbers.

3.7.2 Quartiles and Percentiles

The mean, median, and mode can be thought of as measures of location - they attempt to locate the most representative value. Other measures of location are quartiles and percentiles.

Definition 3.8

Lower, middle, and upper quartiles

The *lower quartile* (q_1) of a set of n measurements x_1 , x_2, ..., x_n which have been ordered from the smallest to the largest is the value of x that exceeds 1/4 , of the measurements and is less than the remaining 3/4 .

The *middle quartile* (q_2) is the median.

The *upper quartile* (q_3) is the value of x that exceeds 3/4, of the measurements is less than the remaining 1/4.

Example 3.10

Find the lower, middle, and upper quartiles for the data set:

20, 34, 17, 18, 28, 33, 12, 1S, 17, 12, 41,
4S, 18, 19, 16, 21, 26, 14, 26, 13, 29

Solution:

Ordered from the smallest to the largest, the 21 measurements are:

12, 12, 13, 14, 15, 16, 17, 17, 18, 18, 19, 20, 21, 26, 26, 28, 29,33,34,41, 45

↑ ↑ ↑

$q_1 = 15.25$ $q_2 = 19$ $q_3 = 28.75$

To determine the first quartile, one fourth of the measurements is 21 /4 = 5.2 and three-fourths is 15.75. We wish to find, therefore, the measurement in the data set such that 5.25 of the measurements are below it and 15.75 are above it. Of course, no such measurement exists, so to find q_1, we must interpolate between the values of the fifth and sixth measurements, 15 and 16. This results in $q_1 = 15.25$.

Definition 3.9

The Pth percentile

The Pth percentile of a set of n measurements $x_1, x_2, \ldots, x_n$ denoted by P, is the value of x such that P percent of the values are less than P and (100 - P) percent of the values are greater than P.

Example 3.11

Find the 85th percentile for the data set in *Example 3.6.*

Solution:

Since 85 percent of n = 21 is 17.85, we are looking for the measurement such that 17.85 of the measurements are below it and 3.15 are above. This value lies between 29 and 33. By interpolation

P = 32.4 .

By looking at the difference among the quartiles, we can get a feel for the variability of the data. One measure of variability using the quartiles is the interquartile range defined by

q3-q1 .

The larger the interquartile range, the more spread out the set of measurements will be.

Tutorial 3.2

1. In the Complaints Department of a large department store on a given day. the lengths (in seconds) of the first 100 telephone calls were recorded (rounded to the nearest second), and the following frequency distribution was constructed:

Class	Class limit	Class boundary	Frequency
1	0-59	-0.5 to 59.5	5
2	60-119	59.5 to 119.5	20
3	120-179	119.5 to 179.5	40
4	180-239	179.5 to 239.5	25
5	240-299	239 5 to 299.5	10

   a. Construct a histogram from this frequency distribution.
   b. Construct a polygon from this frequency distribution.
   c. Compute the approximate mean length of the 100 telephone calls.

2. The following 25 measurements represent the number of business trips taken annually by 25 claim adjusters of the Acme Insurance Company:

   33, 17, 2, 10, 12, 15, 22, 18, 20, 24,8, 27, 8,
   12,17, 15, 21, 38, 16, 18,10,12, 9, 5, 28

   Construct a frequency distribution with 5 classes for this data. Give the relative frequencies and construct a histogram from the frequency distribution. Compute the mean, median, and mode .

3. Find the median ,mean and the mode for the data given in Ex9 ,and Ex10 of tutorial 2.

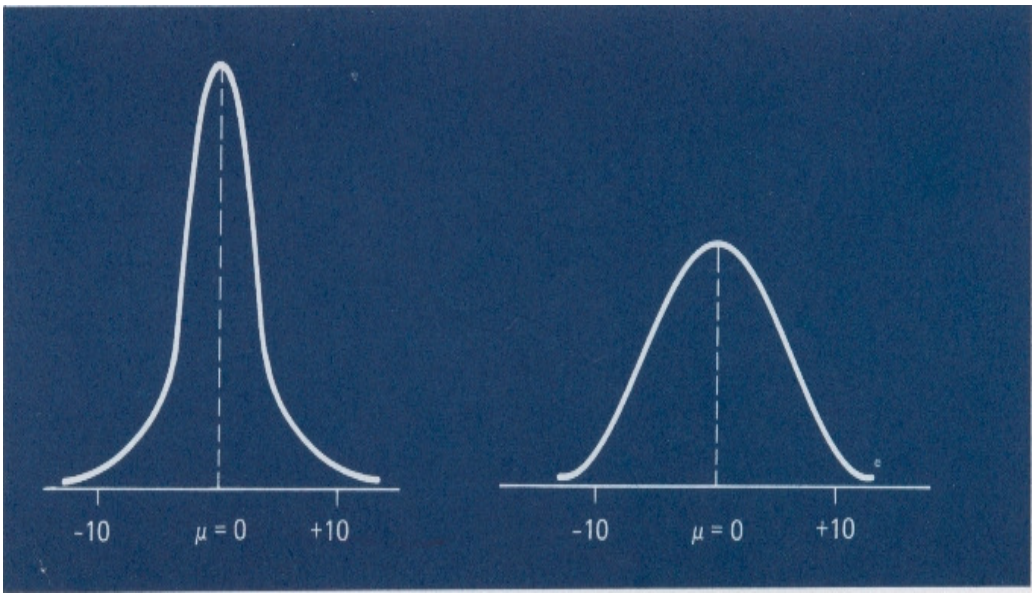
4
Measures of Variation

4.1 Introduction
4.2 The range
4.3 The standard deviation & the variance
4.4 The standard deviation & the variance of grouped data
4.5 Measure of relative variation
4.6 Measure of skewness
Tutorial: 4

4.1 Introduction

While representative measures provide certain information about a distribution, more is needed before a clear picture of the shape of the distribution can be formulated. In figure 4.1 for example, both distributions have the same mean value, but obviously differ in another respect-the amount of dispersion or variability of the values. The concept of variability is very important in statistics. For example, in production management, a major concern is the variability of the quality of a product being produced or the variability of a crucial measurement of a product such as a bearing diameter. More important, in statistical inference, we shall use the concept of variability to determine how good our inferences are. For the moment, it is sufficient to recognize the need for measuring the variability of a set of values to get a better idea of the shape of the distribution of the measurements.

Figure 4.1 Two dissimilar distributions with identical means



4.2 The range

The first and simplest measure of variability is the range.

Definition 4.1

Range

The *range* of a set of measurements $x_1, x_2,..., x_n$ is the algebraic difference between the largest and smallest values.

Example 4.1

Given the following 6 numbers, determine their range.

$x_1=5$, $x_2=0$, $x_3=6$, $x_4=2$, $x_5=-2$, $x_6=9$

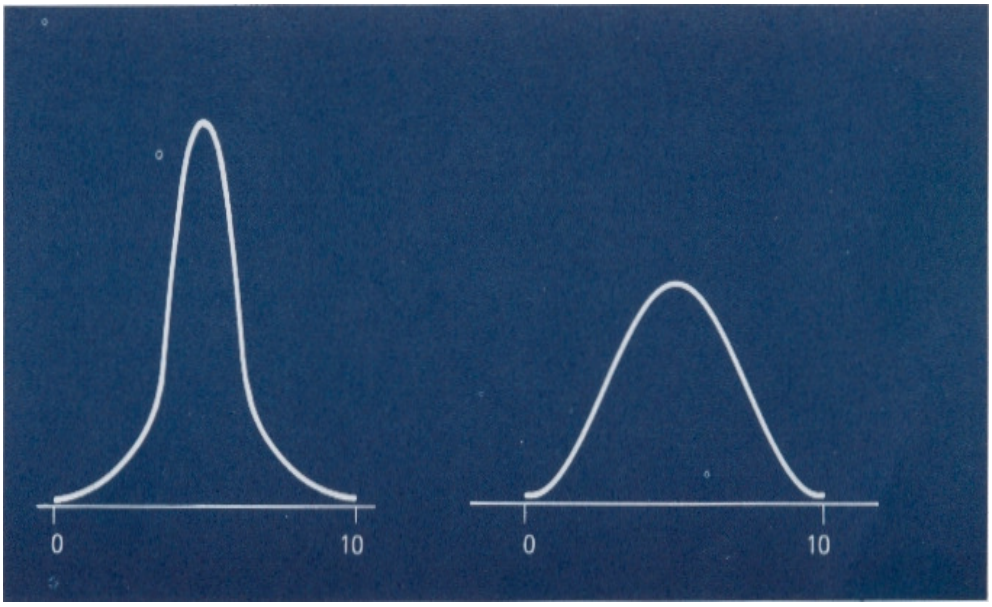
Solution:

The largest number is 9 while the smallest is - 2. Thus the range is 9 -(- 2) = 11.

While the range is very easy to compute, it is not a very satisfactory measure of variability, as figure 4.2 illustrates. The distribution on the left clearly is less variable than the distribution on the right, yet the ranges for the two distributions are identical (10 - 0 = 10).

The mean of these six measurements is:

Figure 4.2 Two distributions with equal ranges



4.3 The Standard Deviation and the Variance.

Since the variation of a set of numbers is small if they are bunched closely about their mean and it is large if they are spread over considerable distances away from their mean, it would seem reasonable to define variation in terms of the distances (deviations) by which numbers depart, from their mean. If we have a set of numbers x_1, x_2,...,and x_n whose mean is $\bar{x}$, we can write the amounts by which they differ from their mean as $\bar{x}_1 - x$, $\bar{x}_2 - x$,..., and $\bar{x}_n - x$. These quantities are called the deviations from the mean and it suggests itself that we might use their average, namely, their mean, as a measure of the variation of the numbers. This would not be a bad idea, if it were not for the fact that we would always get 0 for an answer, no matter how widely dispersed the data might be. As the quantity $\sum (x_i - \bar{x})$ is always equal to zero-some of the deviations are positive, some are negative, but they "average out," that is, their sum as well as their mean are always equal to zero.

Since we are really interested in the magnitude of tile deviations and not in their signs, we might simply "ignore" the signs and, thus, define a measure of variation in terms of the absolute values of the deviations from the mean. Indeed, if we added the values of the deviations from the mean as

if they were all positive and divided by n., we would obtain a measure of variation called the mean deviation (see Exercises 5 and 6 on tutorial V). Unfortunately, this measure of variation has the drawback that, owing to the absolute values, it is difficult to subject it to any sort of theoretical treatment; for instance, it is difficult to study mathematically how in problems of sampling, mean deviations are affected by chance. However, there exists another way of eliminating the signs of the deviations from the mean, which is preferable on theoretical grounds: The squares of the deviations from the mean cannot be negative; in fact, they are positive unless a value happens to coincide with the, mean, in which case $\sum(x_i-x)$is equal to zero.

Definition 4.2

Variance

The *variance* of n measurements $x_1,x_2,\ldots,x_n$ is denoted by s^2 , is given by

$$s^2 = \sum(x_i - \bar{x})^2/n,$$

and this is how, traditionally, the variance has been defined. Expressing literally what we have done here mathematically, it has also been called the mean-square deviation.

Nowadays, it has become the custom among most statisticians and research workers to make a slight modification in this definition, which consists of dividing the sum of the squared deviations from the mean by n- 1 instead of n. Following this practice, which will be explained later, let us thus formally define the sample variance , as

$$s^{*2} = \sum(x_i - \bar{x})^2 /(n-1) .$$

Definition 4.3

Standard deviation

The *standard deviation* is denoted by s(s*) is given by

$$s= (\sum(x_i - \bar{x})^2 / n)^{1/2} ,$$

or

$$s^*= (\sum(x_i - \bar{x})^2 /(n-1))^{1/2} .$$

The formulas we have given so far in this section are meant to apply to samples, but if we substitute μ for $\bar{x}$ and N(or N-1) for n(or n-1), we obtain analogous formulas for the standard deviation and the variance of a population. It has become fairly general practice to write population standard deviations as S when dividing by N and S* when dividing by N - 1; symbolically,

Definition 4.4

Population Standard Deviation

The *Population standard deviation* of N observation is denoted by S(S*) is given by

$$S = \left(\sum (x_i - \mu)^2 / N\right)^{1/2},$$

or

$$S^* = \left(\sum (x_i - \mu)^2 / (N-1)\right)^{1/2}.$$

To explain why we divide by n- 1 instead of n and N - I instead of N in the formulas for s and S, let us point out that if we wanted to use sample variances to estimate the respective variances of the populations from which the samples were obtained, division by n instead of n - 1 would give us values which on the average are too small. We cannot prove the following at the level of this note, but it is shown in most textbooks on mathematical statistics that the values would be too small on the average by the factor $(n-1)/n$.

For instance, for n = 5 the estimates would on the average be (5-1)/5 = 0.80 or 80 per cent of what they should be,and hence 20 per cent too small. To compensate for this we divide by n - 1 instead of n in the formulas for the sample standard deviation and the sample variance. As the statisticians say, this makes the sample variance s^2 unbiased; that is, if we calculate s^{*2} for several samples taken from the same population, the values we get should average S^2, the variance of the population. Note, however, that this modification is of no significance unless n is small; generally, its effect is negligible when it is large, say 100 or more. The same applies to the difference between S^2 and S^{*2}, which is negligible unless the size of the population is very small, and in actual practice this is usually not the case.

Example 4.2

To illustrate the calculation of a sample standard deviation, let us find s for the following data on the number of burglaries reported in a town during the first six weeks of 1972: 12, 18, 7, 11, 15, and 9. First calculating x, we get

$$\bar{x}=(12+18+7+11+15)/6=12.$$

and then the remainder of the calculations are as shown in the following table

x	**$(x-\bar{x})$**	**$(x-\bar{x})^2$**
12	**0**	**0**
18	**6**	**36**
7	**-5**	**25**
11	**-1**	**1**
15	**3**	**9**
9	**-3**	**9**
	0	**80**

and

$$s^*= \left(\sum(x_i - \bar{x})^2 /(n-1)\right)^{1/2}$$

$$=(80/(6-1))^{1/2}=(16)^{1/2}=4 \cdot$$

Thus, $\bar{x}$ = 12 provides us with an estimate of the average number of burglaries in this town per week, and the value of the standard deviation,

s* = 4, tells us something about the variability of the figures from week to week. How such a value of s is to be interpreted will be discussed

and how it can be used to judge how close $\bar{x}$ = 12 might be to μ, the true average number of burglaries in this town per week, will be discussed. (Note that in the above table we totaled, as a check, the $\bar{x}$ - x column, and as we have indicated , the sum of its values must always equal zero.)

The calculation of s was very easy in this example, and this was due largely to the fact that the x's, their mean, and hence also the deviations from

the mean were all whole numbers. Had this not been the case, it might have been profitable to use the following short-cut formula for s:

$$s^*=((n(\sum x^2) - (\sum x)^2)/n(n-1))^{1/2}.$$

This formula does not involve any approximations and it can be derived from the other formula for s by using the rules for summations. The advantage of this short-cut formula is that we do not have to go through the process of actually finding the deviations from the mean; instead we calculate $\sum x$, the sum of the x's , $\sum x^2$, the sum of their squares, and substitute directly into the formula. Referring again to the burglary data, we now have

x	x^2
12	144
18	324
7	49
11	121
15	225
9	81
72	944

$$s^*=((n(\sum x^2) - (\sum x)^2)/n(n-1))^{1/2},$$

$$= (6(944)-(72)^2/6.5)^{1/2} \quad =4.$$

It appears that in this particular example the "short-cut" method is actually more involved; this may be the case, but in actual practice, when we are and dealing with realistically complex data, the short-cut formula usually provides considerable simplifications.

Example 4.3

To demonstrate the advantages of the short-cut formula, let us determine the sample variance of the numbers 12, 7, 9, 5, 4, 8, 17, 2, 11, 14, 13, and 9, using first the formula definition 4.3 and then the short-cut formula. Without using the short-cut formula we get

x	$(x- \bar{x})$	$(x- \bar{x})^2$
12	2.75	7.5625
7	-2.25	5.0625
9	-0.25	0.0625
5	-4.25	18.0625
4	-5.25	27.5625
8	-1.25	1.5625
17	7.75	60.0625
2	-7.75	52.5625
11	1.75	3.0625
14	4.75	22.5625
13	3.75	14.0625
9	-0.25	0.0625
111	0	212.2500

and

$\bar{x} = (111/12)=9.25$,

$s^{*2} =(212.2500/11)=19.3$,

and working with the short-cut formula we get

x	x^2
12	144
7	49
9	81
5	25
4	16
8	64
17	289
2	4
11	121
14	196
13	169
9	81
111	1239

$$s^{*2}=(12(1239)-(111)^2)/12.11$$
$$=19.3.$$

Here the short-cut formula provided considerable simplifications.

A further simplification in the calculation of s* or s^{*2} consists of adding all arbitrary positive or negative number to each measurement. It is easy to prove that this would have no effect on the final result, and had we used this trick in the last example, we might, have subtracted 10 (added -10) from each number, getting 2, -3, -1, -5, -6, -2, 7, -8, 1, 4, 3, and – 1 instead of the original numbers. The sum of these numbers is -9, the sum of their squares is 219, and substitution into the formula for
s^{*2} yields

$$s^{*2}=(12(219)-(-9)^2)/12.11$$

$$=19.13.$$

which is exactly what we had before. Since the purpose of this trick is to reduce the size of the numbers with which we have to work, it is usually desirable to subtract, a number that is close to the mean. In our example the mean was 9.25, and the calculations might have been even simpler if we had subtracted 9 instead of 10. Although the short-cut formula was given for use with samples, we have only to substitute N for n throughout to make the formula applicable to the calculation of s^{*2} or s*.

4.4 Standard deviation and Variance for grouped data.

If we want to calculate the standard deviation of data which have already been grouped, we are faced with the same problem as on the mean. Proceeding as we did in connection with the mean, and assigning the value of the class mark to each value falling into a given class.

Definition 4.5

Standard deviation for Grouped Data

The *standard deviation* of grouped data is

$$s^* = \left(\left(\sum(x_i - \overline{x})^2 \cdot f_i/(n-1) \right)^{1/2} \right.,$$

$$n=\sum fi,$$

and, if we substitute μ for $\bar{x}$ and N or N - 1 for n - 1, we obtain analogous formulas for s. Note that in this formula the x's are now the class marks and the f's are the corresponding class frequencies.
The above formula serves to define s for grouped data, but it is seldom used in actual practice. Either we use a computing formula analogous to the short-cut formula ,

$$s^*=((n(\sum(xi^2\ fi)-(\sum xifi)^2)/n(n-1))^{½},$$

where the x's are the class marks and the f's the corresponding class frequencies, or we use the same kind of coding as in the calculation of the mean of grouped data.

Following of the mean, we obtain

$$s^*=c((n(\sum(ui^2\ fi)-(\sum uifi)^2)/n(n-1))^{/12}$$

This is the coding formula for computing the standard deviation of grouped data. Note that this formula can be used only when the class intervals are all equal.
Although this short-cut formula may look fairly complicated, it makes the calculation of s very easy. Instead of having to work with the actual class marks arid the deviations from the mean, we have only to find the sum of the products obtained by multiplying each u by the corresponding f, the sum of the products obtained by multiplying the square of each u by the corresponding f, and substitute into the formula.

Example 4.4
To illustrate the use of this short-cut formula for the calculation of s for grouped data, let us refer again to the distribution of the scores of the150 applicants. Using the same u-scale, we get

Class Marks xi	ui	fi	ui .fi	$ui^2.fi$
14.5	**-4**	**1**	**-4**	**16**
24.5	**-3**	**6**	**-18**	**54**
34.5	**-2**	**9**	**-18**	**36**
44.5	**-1**	**31**	**-31**	**31**
54.5	**0**	**42**	**0**	**0**
64.5	**1**	**32**	**32**	**32**
74.5	**2**	**17**	**34**	**68**
84.5	**3**	**10**	**30**	**90**
94.5	**4**	**2**	**8**	**32**
			33	**359**

and

$$s^*=c((n(\sum(ui^2\, fi)-(\sum uifi)^2)/n(n-1))^{/12},$$

$$=10((150(359)-(33)^2)150.149)^{½}$$
$$=15.4\ .$$

The variation of the scores of the 150 applicants is, thus, measured by a standard deviation of 15.4, and we shall indicate below how such a figure might be interpreted.

4.5 Measures of Relative Variation

The standard deviation of a set of measurements is often used as an indication of their inherent precision. If we repeatedly measure the same quantity, for example, a person's temperature, the mileage a person gets with a certain gasoline, or the weight of a piece of rock brought down from the moon, we would hardly expect always to get the same result. Consequently, the amount of variation we do find in repeated measurements of the same kind provides us with information about their precision. To give an example, suppose that 5 measurements of the length of a certain object have a standard deviation of 0.20 in. Although this information may be important, it does not allow us to judge the relative precision of these measurements; for this purpose we would also have to know something about the actual size of the quantity we are trying to measure. Clearly, a standard deviation of 0.20 in. would indicate that the measurements are very precise if we measured the span of a bridge; on the other hand, they would be far from precise if we measured the diameter of a small ball bearing.

This illustrates the need for measures of relative variation, that is, measures which express the magnitude of the variation relative to the size of whatever

is being measured. The most widely used measure of relative variation is the coefficient of variation.

Definition 4.6
Coefficient of Variation

The *coefficient of variation* of grouped (or non-grouped data) is denoted by CV ,is given by the formula

$$CV = (s/\bar{x}) \cdot 100.$$

This simply expresses the standard deviation of a set of data (or distribution) as a percentage of its mean. When dealing with populations, we ,analogously define the coefficient of variation as

$$CV = (S/\mu) \cdot 100$$

If in the above example the standard deviation s = 15.4 and $\bar{x}=56.7$,then

$$CV=(15.4/56.7) \cdot 100.$$

By using the coefficient of variation, it is also possible to compare the dispersions of two or more sets of data that are given in different units of measurement. Instead of having to compare, say, the variability of weights in pounds, lengths in inches, ages in years, and prices in dollars, we can instead compare the respective coefficients of variation-they are all percentages.

4.6 Skewness Measure

As suggested earlier in this chapter, one possible measure of the skewness of a distribution of a set of measurements is the difference between its mean and its median. We will use a function of this difference as a measure of skewness.

Definition 4.7
Skewness measure(Sk)

The *skewness measure* of a set of n measurements $x_1,x_2,...,x_n$ (or grouped data)denoted by Sk, is defined as three times the difference between the mean and the median, divided by the standard deviation:

$$Sk=3(\bar{x} - m)/s$$

If the distribution is skewed right, the mean will be larger than the median and Sk will be positive. If the distribution is skewed left, the mean will be smaller than the median and Sk will be negative.
The effect of dividing by s in Sk is to produce a statistic which is not dependent on the unit of measurement. The mean, median and standard deviation are all measured in the same units for a given data set.

The ***skewness measure*** Sk can be used in two ways. First, the sign of Sk indicates the direction of skewness: +, skewed right and -, skewed left. Second, if Sk is larger in magnitude in one data set than in another, the first data set distribution is more skewed than the other. That is, Sk can be used as a relative measure of the degree of skewness among data sets.

Example 4.5
Compute Sk for the following set of 5 measurements:

$$x_1=10 , x_2=4 , x_3=4 , x_4=6 \text{ and } x_5=1.$$

Solution:

The mean, median and standard deviation for this data set are $\bar{x} = 5$, $m = 4$ and $s= 2.97$. Therefore,

$$Sk=3 . (5-4)/2.97=1.01$$

Since Sk is positive, the distribution of the 5 measurements is skewed to the right.

Tutorial 4

1. Find the range , s^2, s ,CV &Sk of the data of Ex. 6,7 & 8 of tutorial 3.
2. The following is the distribution of the percentage of students belonging to a certain minority group in 40 schools:

Percentage	Frequenc
0- 4	14
5- 9	11
10-14	7
15-19	6
20-24	2

a. Calculate the mean.
b. Calculate the mode.
c. Calculate the median.
d. Calculate s^2 for this distribution
   (i) without coding;
   (ii) with coding.
e .Calculate C.V.
f. Calculate Sk.

3.Consider the following frequency distribution

Class	Freq.
0-2	10
3-5	6
6-8	3
9-11	1

Calculate :
a.s^2 & s.
b.median.
c.mode.
d. CV.
e. Sk.

4.Suppose that the random variable x has the following table

Class	Freq.
0-7	2
8-15	10
16-23	8
24-31	3
32-39	2

a. Calculate the mean.
b. Calculate the mode.
c. Calculate the median.
d. s^2 & s .
e. C.V .
f. range.
j. Sk.

5

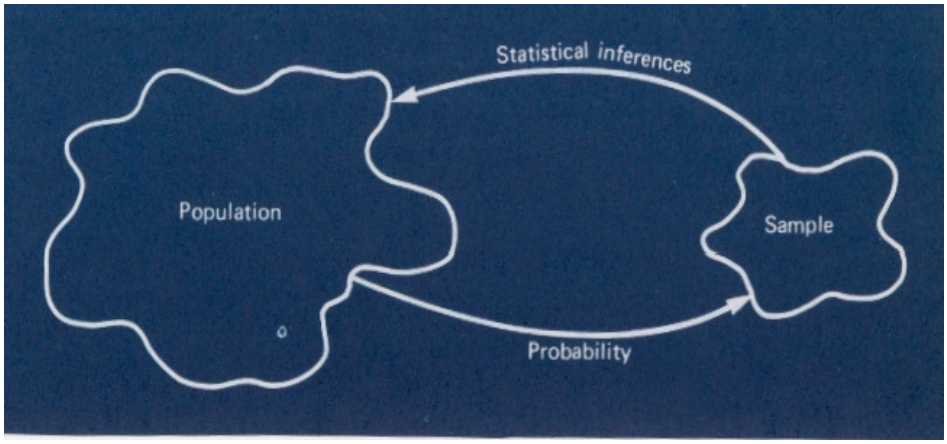
Introduction to Probability & Random Variables

5.1 Introduction
5.2 The sample and event spaces
5.3 Computing probabilities from the sample space
5.4 Permutations, combinations, and other counting rules
5.5 Random variable
5.6 Probability mass function
5.7 Probability density function
Tutorials 5.1 & 5.2

5.1 Introduction

Probability plays an integral role in inferential statistics by building a "bridge" between the population and the sample taken from it. Our initial applications of probability in this connection will be to make deductions about a sample from a known population .The use of probability is as indicated in figure 5.1: probability reasons from the population to the sample, while statistical inferences are drawn about the population from the sample.

Figure 5.1 Role of probability in the statistical inference process



Example 5.1

As an example of the use of probability in this context, consider the national election for the office of president of the United States. Let us suppose that only two candidates are listed on the ballot for the presidency, the Democratic candidate (A), and the Republican candidate (B). Further, suppose it is known in the population of registered persons who will vote on election day that 60 percent will vote for A and 40 percent will vote for B. If we now randomly sample one person from this population, what is the probability that he or she will vote for A? Since we know that 60 percent of the persons will vote for A, the probability that the one sampled person will vote for A is 0.60. Knowledge of the probabilities of the two possible outcomes of the experiment (voting for A or voting for B) enables us to deduce the probability of the outcome in our sample of one.

Indeed, by using this knowledge, we could deduce the probabilities of zero, one, or two persons voting for A in a sample of two, and so on for

larger sample sizes. Thus, if the population is known in the sense that the probabilities associated with the values in it are known, then this knowledge can be used to deduce the probabilities of the outcomes in the sample.

To illustrate the use of probability in making inferences from a sample to the population, suppose candidate A conjectures before the election that 60 percent of the people in the voting population will vote for him. In order to check this conjecture, his campaign manager randomly samples ten individuals from the population and finds that all ten intend to vote for candidate B. If the probability of a randomly chosen person voting for A is really 0.60, it is extremely unlikely that ten randomly chosen persons would all vote for candidate B. It is more likely that the true percentage who will vote for A is something considerably less than 60 percent. Hence, knowledge of this experiment (sample outcome) indicates to A that more resources (campaigning, etc..) may have to be employed if he is to have a chance of winning the election. Candidate A is interested, of course, in testing whether this sample is indicative of the population characteristics (voting patterns), whether more sample information should be obtained, or whether the election is likely to go to B (in which case A would be wasting his time by campaigning further).

In practical situations, probability is used as a vehicle in drawing inferences about unknown population characteristics. Additionally, as we shall see later, probability concepts can be used to give us an indication of how good these inferences are.

In this chapter, we will assume the population is known and compute the probability of the occurrence of various sample outcomes. In effect, we will be selecting a probability model depicting the outcomes in the population. In practical applications of statistics, we shall see that the selection of this model is an integral part of the statistical inference process.

5.2 The Sample and Event Spaces

In the presidential election example discussed in the previous section, we defined a population which consists of registered persons who will vote on election day. Suppose we assign a"1" to an individual if he or she intends to vote for candidate A, and "0" if he or she intends to vote for candidate B. The population can then be thought of as a collection of ones and zeroes. How are these ones and zeroes generated? Each person in the population must be contacted and represented by a"1" or a "0." The process of con-

tacting each person to determine the outcome ("1" or "0") is called an experiment.

Definition 5.1
Experiment
An *experiment* is a process which results in one and only one outcome of a set of disjoint outcomes, where the outcomes cannot be predicted with certainty.

In the voting example, there are only two possible outcomes, and they are disjoint (non-overlapping): a zero and a one. With our previous assumption that only two candidates are listed on the ballot, each experiment results in one and only one of the two possible experimental outcomes. And we cannot predict with certainty the outcome before the experiment is conducted. Repeated trials of this experiment will generate the population of zeroes and ones. Other examples of experiments are:

Example 5.2

A professor at a large university is selected and his salary is recorded.

Example 5.3
A unit of a product is selected from an assembly line and is analyzed to determine whether it is defective.

Example 5.4
A light bulb is randomly selected from the day's production and its time to failure measured.

By repeating an experiment many times, a population of outcomes can be generated. For example, if we repeated the experiment in 3. until each and every light bulb in the day's production run had been tested to failure, the population of all times to failure of this set of light bulbs would have been generated. In the process of doing this, it should also be noted that the entire day's production of light bulbs (the population) would have been destroyed. We can also think of the sample being generated by repeated trials of an experiment. For example, if we wanted to sample ten light bulbs, we could repeat the experiment ten times.
The outcomes of an experiment are called simple events. Simple events shall be denoted by the capital letter E subscripted to associate E; with a particular outcome (ith) of an experiment.

Definition 5.2
Simple event
A *simple event* is the outcome of an experiment.

Example 5.5
Suppose in our presidential election example that we randomly sample two persons in the population of voters. A possible set of simple events associated with this experiment is:

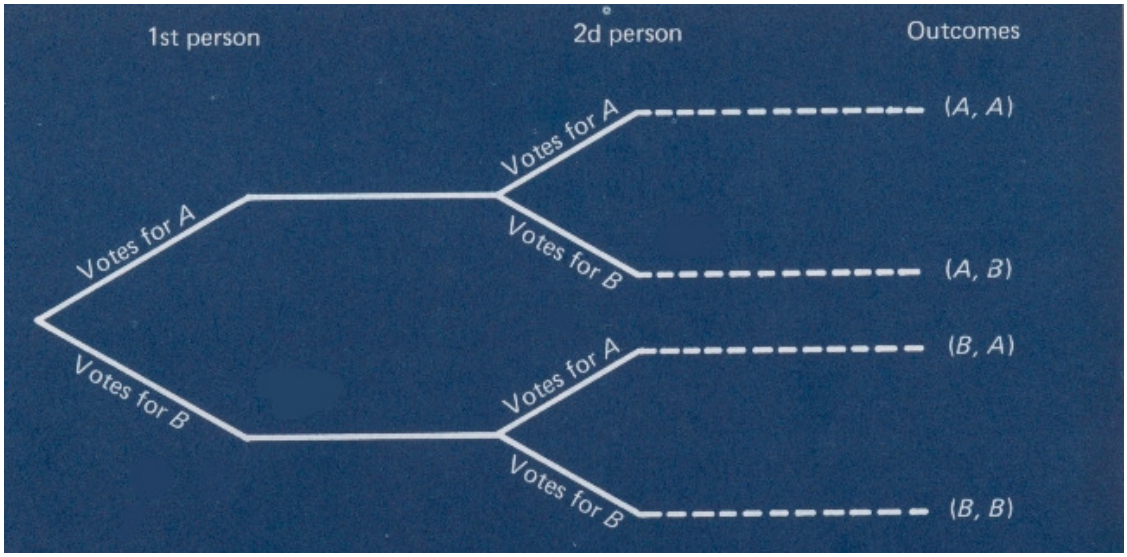
Vote By

Event	First person	Second person
E 1	**A**	**A**
E 2	**A**	**B**
E 3	**B**	**A**
E 4	**B**	**B**

These outcomes of the experiment can also be indicated by using an outcome tree as shown in figure 5.2. The first "branch" of the tree is associated with the first person who votes either for A or for B. The second set of branches is associated with the second person who votes either for A or for B also.

The outcome tree represents a logical way to list the simple events of an experiment. It is very practical if the number of events is not too large.

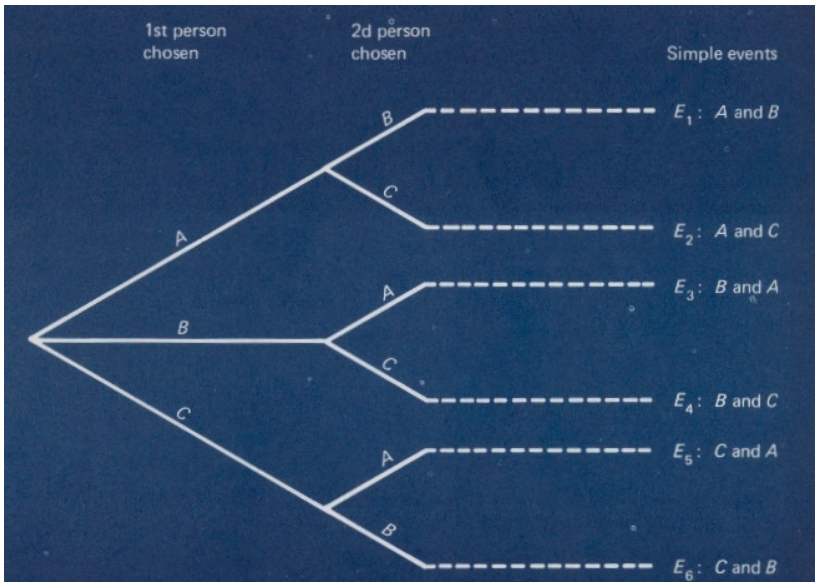
Figure 5.2 Outcome tree for *Example 5.5*



Example 5.6

Suppose three persons, A, B, and C, are interviewing for a job. Two will be hired. The experiment is the selection of two of the three interviewed individuals for the job. The simple events can be listed using an outcome tree as illustrated in figure 5.3.

Figure 5.3 Outcome tree for *Example 5.6*



Notice in *Example 5.6* that the six simple events listed specify not only the two individuals selected, but also the order in which they are selected. That is, E, and E, both result in the first two individuals, A and B, being selected. If the order in which the two individuals are selected is not important, then we need not distinguish between E_1 and E_2, E_3 and E_5. and E_5, and E_6. In this case, a simpler set of outcomes would be:

E_1*: A and B are selected,

E_2*: A and C are selected,

E_3*: B and C are selected.

As suggested above, it is often possible to define the outcomes and the experimental simple events differently in the same experiment. To gain an

understanding of how to define the simple outcomes of an experiment, consider the following example.

Example 5.7

Assume that Herman is to toss a "fair coin" twice. He informs you that there are three possible outcomes (simple events) of this experiment:

E_1*: No heads (two tails),
E_2*: One head (one tail),
E_3*: Two heads (no tails).

Herman tells you that the probability of any one of the three simple events occurring is 1/3.He then wishes to wager with you on the outcome of one trial of the experiment, say E_2-one head occurring in two tosses of the coin.

Before deciding to accept a wager, you construct an outcome tree of a single trial of the experiment.

From the outcome tree, it is clear that we may define another set of simple events for this experiment:

E_1: (H,H),

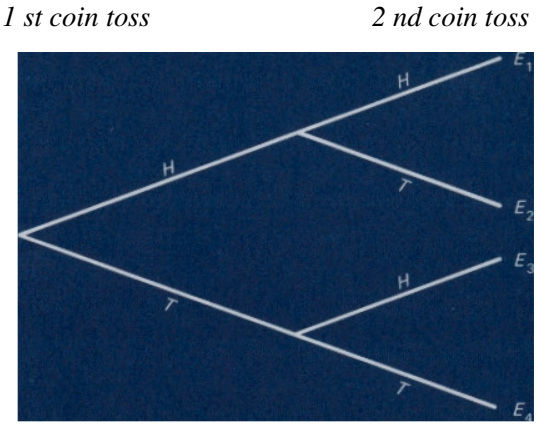
E_2: (H.T),
E_3: (T,H),
E_4: (T,T).

If the coin is "fair," then the probability of each of the outcomes E_1, E_2, E_3 and E_4 in figure 5.4 occurring is 1/4.

In terms of the original three outcomes, E^*_1, E^*_2 and E^*_3 it is clear that each does not have a 1/3, probability of occurring - the proper probabilities are:

$P(E_1^*) = 1/4$, [$E_1^* = E_1$ (H,H)],
$P(E_2^*) = 2/4$, [$E_2^* = E_2$ or E_3 (H,T) or (T,H)],
$P(E_3^*) = 1/4$, [$E_3^* = E_4$ (T,T)].

Figure 5.4 Outcome tree for a coin-tossing experiment



Definition 5.3

Event

An *event* is a subset of outcomes of an experiment.

Notice that any simple event of an experiment is an event because it is a single outcome of the experiment.

Definition 5.4

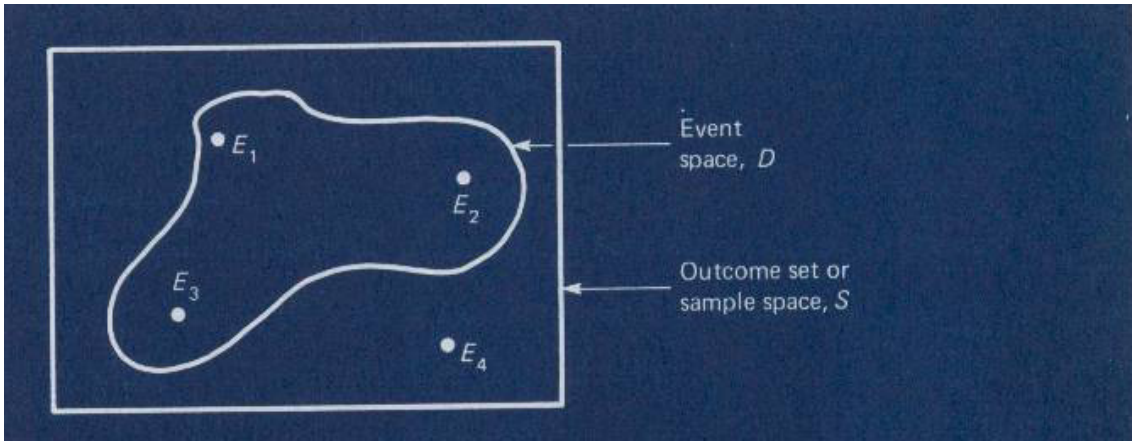
Null event

A *null event* is an event containing no simple events in an experiment. It is denoted by : ϕ.

In *Example 5.5*, an example of a null event is "no persons vote for A or for B." It is impossible for this event to happen, because there are only two candidates, A and B, and the population consists of individuals who will vote in the election. In this instance, the event set is empty, for it does not contain any of the simple events in the experiment.

The simple events of an experiment and events defined to be collections of these simple events can be portrayed graphically by a Venn diagram. The Venn diagram associated with the simple events in *Example 5.5 and* the event D defined above is shown in figure 5.5. Each simple event in a

Figure 5.5 A Venn diagram for the simple events in *Example 5.5*



Venn diagram is shown as a "point" with its corresponding subscripted letter E. The collection of all simple events in an experiment is the complete set of sample points in the Venn diagram and is called the sample space. The event D is illustrated in the Venn diagram by enclosing the sample points belonging to it; the resulting closed region is called the event space, D.

Definition 5.5

Sample point

A *sample point* is a simple event in an experiment.

Definition 5.6

Sample space

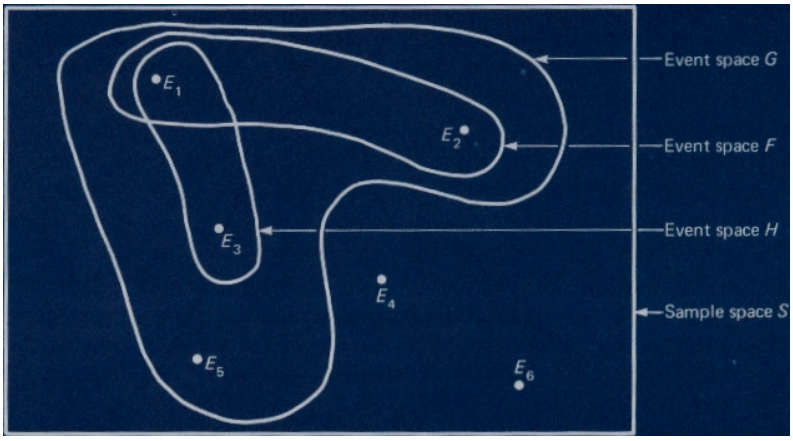
A *sample space* is the set of all possible outcomes of an experiment.

Definition 5.7

Event space

An *event space* is the collection of sample points corresponding to an event defined over the sample space.

Figure 5.6 Venn diagram for the experiment shown in Figure 5.3



In *Example 5.6* suppose that we define the following events:

F: A is chosen first

G: A is chosen without regard to selection order

H: A and B are chosen without regard to selection order.

Assume that we define the simple events as in Figure 5.3:

E1: A and B E3: B and A E5: C and A

E2:A andC E4:B andC E6: C and B.

The Venn diagram showing the sample space for this experiment, the sample points, and the event spaces F, G, and H is illustrated in figure 5.6.

5.3 Computing Probabilities from the Sample Space

In Example 5.5, where two persons are randomly selected from the voting population, there are four possible outcomes of the experiment. In the corresponding sample space for this experiment, illustrated in Figure 5.5, we defined the event D to be, "at least one of the two persons votes for candidate A." What is the probability that the event D occurs in this experiment? This question can be answered directly from the sample space associated with the experiment if the probabilities of the simple events occurring are known. Thus, to answer a probability question about an event in an experiment, we first must assign probabilities to the simple events

associated with the experiment. We will denote by $P(E_i)$ the probability assigned to the simple event E_i. The assigned probabilities $P(E_1)$, $P(E_2)$, . . . , $P(E_m)$, where there are M simple events in the experiment sample space, must satisfy three probability axioms for experiments that have a finite number of outcomes.

Axiom 1

$$0 \leq P(E_i) \leq 1 \text{ for } i = 1, 2, ..., m.$$

The *first axiom* requires that every simple event be assigned as its probability a non-negative number between 0 and I inclusive.

Axiom 2

$$\sum_{i=1}^{m} P(E_i) = 1 \ .$$

The *second axiom* requires that the probabilities assigned to all the simple events in the experiment must total one.

Axiom 3

$$P(E_1 \text{ or } E_2 \text{ or } E_3 \text{ or } ...) = P(E_1) + P(E_2) + P(E_3) + \ldots$$

The *third axiom* requires that the probability of one or more members of a set of simple events occurring in an experiment is the sum of their respective probabilities.

Example 5.8
Suppose in *Example 5.6* we assign the following probabilities to the simple events in that experiment:

	Simple events	**Probability**
E_1	**(A,A)**	**$P(E_1) = 1/4$**
E_2	**(A,B)**	**$P(E_2) = 1/4$**
E_3	**(B,A)**	**$P(E_3) = 1/4$**
E_4	**(B,B)**	**$P(E_4) = 1/4$**

These probability assignments satisfy the above three axioms-each probability assigned is a positive number between 0 and I inclusive, the probabilities total one, and the probability that any member of a collection of the simple events will occur is the sum of the probabilities of the members in the collection.

These axioms are intuitive; most of us have an understanding of them before taking any formal training in probability. If an event is certain to happen, its probability of occurrence is 1 and if an event is certain not to happen, its probability of occurrence is 0.
How do we, in fact, formally define probability? We will consider one way of defining probability.

Definition 5.8

Relative frequency definition of probability

If an event E is defined in an experiment, the experiment is repeated a very large number of times, say N, and the event E is observed to occur in n of these N experimental trials, then

$$P(E) = n/N .$$

The ratio n/N represents the proportion of the time that event E occurs in repeated experiments.

5.4 Permutations Combinations and other Counting Rules

Numerous counting rules can be used to count the number of points in sample and event spaces. We shall consider four of the most important counting rules. Each will be presented without proof and followed by examples.

Rule5.1

m.n rule

Suppose that there are m distinguishable objects in one group and n dis-

tinguishable objects in another group. If one element is selected from each group, it is possible to form m . n pairs of objects.

Example 5.9

Herman has decided to purchase a new hi-fi system with the money he saved by buying a compact car instead of a large sedan. His hi-fi system will be composed of a receiver, a pair of speakers, a record changer, and a tape deck. In the store where he will make the purchase, there are 10 different kinds of receivers, 5 kinds of speakers, 4 kinds of changers, and 8 kinds of tape decks. How many systems can Herman choose from?

Solution:

Since he must select one element from each of the four groups, he can choose from (10)(5)(4)(8) = 1600 possible systems.

The next two counting rules apply to a different type of experiment as indicated in the following example.

Example 5.10

Suppose three persons, A, B, and C, are competing for two job positions. How many ways can two people be selected for employment from the three?

Solution:

In this problem, it is easy to list the possible different outcomes of the experiment. There are three: AB, AC, and BC. However, we may be interested in the order of the selection as well as the content of the resulting pairs. If this is the case, there are six possible outcomes: AB, BA, AC, CA, BC, and CB.

In Example 5.10, there are two different ways to view the pairs formed by selecting two persons out of three. Suppose there are n distinguishable objects from which we are selecting a subset of size r. If we are concerned about the number of groups of size r that can be formed from the n where one group of size r is different from another if its content is different, we want to determine the number of combinations of r things selected from n. If, on the other hand, we want to compute the number of groups of size r that can be formed from n where one group of size r is different from another in terms of both its content and the order in which the r things were drawn, then we want to determine the number of permutations of r things drawn from n.

In Example 5.10, the number of combinations of two persons drawn

from three is 3, while the number of permutations of two persons drawn from three is 6.
In conjunction with the rules for computing the number of permutations and combinations, the complete definitions follow.

Definition 5.9
Permutations
An ordered arrangement of r distinguishable objects is called a *permutation*. The number of *permutations* of r objects taken from n distinguishable objects will be denoted by P_r^n.

Definition 5.10
Combinations
A set of r distinguishable objects is called a *combination*. The number of *combinations* of r objects taken from n distinguishable objects will be denoted by C_r^n.

Rule 5.2
Permutations

$$P_r^n = n!/(n-r)!$$

Rule 5.3

Combinations

$$C_r^n = n!/r!(n-r)! = (1/r!)\, P_r^n$$

The symbol n! is called "n-factorial"; n! = n(n - 1)(n - 2) ...(2)(1). Thus,

4! = 4(3)(2)(1) = 24 and 6! =(6)(5)(4)(3)(2)(l) = 720; 1! = 1 and, by definition, 0! = 1.

Example 5.11

A committee of three is to monitor the activities of the local club. The committee is to be formed by selecting three people from a group of five persons. How many different committees could be formed?

Solution:

Nothing is mentioned about the order or arrangement of the three selected individuals. Thus, one committee will be different from another if it has one

or more different people in it. We are only concerned about the content of each group, and therefore we want to determine the number of combinations of three things taken from five.

$$C_3^5 = 5!/ 3!(5 - 3)!=5.4.3.2.1/(3.2.1)(2.1)$$
$$=10,$$

thus, it is possible to develop 10 different committees of three people ;elected from five. You can check this result by listing all possible groups of three drawn from five, where the five people are labeled A. B, C. D, and E.

Example 5.12
A club committee of three is to be formed by selecting three people from a group of five. One of the selected people will be chosen a chairman of the committee, another the secretary, and the third person will simply be a "member" of the committee. How many different committees can be formed?

Solution:
Suppose the three people (denoted by A, B, and C, respectively) have been chosen from among the five. Once we have this combination of three people, we must then assign them to the three positions: chairman, secretary, and member.

We can view this process as ordering the three persons. That is, let the first position in the ordering (or permutation) be the chairman, the second be the secretary, and the third be the member. The possible permutations are: ABC, ACB, BAC, BCA, CAB, and CBA. Each of these combinations is a different committee, although each contains the same three people, A, B, and C. Thus, we want to compute the number of permutations of r = 3 people chosen from n = 5.

$$P_3^5=5!/(5-3)!=60$$

In this case, there are 60 different committees which can be formed. Notice that this is six times as many committees that could be formed in Example 5.11 because for each combination in Example 5.11, there are 6 permutations in Example 5.12.

Tutorial 5.1

l. Give the sample space of each of the following experiments in the form of a Venn diagram. Be certain to define the simple events corresponding to the sample points in the sample space.
   a. A fair coin is tossed three times.
   b. A coin and a die are tossed together.
   c. Two fair dice are tossed, and the sum of the dots of the two faces turning up is recorded.
   d . A student receives his score on a multiple choice exam containing 20 questions.
   e. A student receives his grade on an exam.
   f. The number of telephone calls received at a switchboard during a live minute interval is recorded.
   g. A child is selected in a first grade class, and his or her weight (to the nearest pound) is recorded.

2. A committee is composed of two men and two women. One member of the committee is selected to serve as chairman and another is selected to serve as secretary.
   a. Define the simple events comprising the sample space of this experiment. Identify which sample points in the sample space belong to the following event spaces:
   b. The younger man is selected as chairman: Event A.
   c. A man is selected as chairman: Event B.
   d. A woman is selected as secretary: Event C.
   e. Events A and C occur: Event D.
   f. Events B or C or both occur: Event E.
   g. Show the live event spaces in a Venn diagram.

3. Two college job recruiting officers. Herman and Bill, come to the University of Truth campus to fill positions in their organizations. Each officer is attempting to fill three positions. Three students qualify for the positions described. and each will be interviewed by the two officers. If a sample point is defined as a specific number of students hired by Herman or Bill, define the following events as specific collections of sample points: Note: there are six jobs and three students three jobs will not be filled. (Hint: The sample space is two-dimensional.)
   a. The sample space S which consists of all outcomes defining the number of students hired by each officer.

b. Event A: Herman hires at least two students.

c. Event B: All three students are hired by Bill.
d. Event C: Exactly one student is hired by Bill.
e. Event D: Bill hires two students and so does Herman.

4. When one card is drawn from a well-shuffled deck of 52 playing cards what are the probabilities of getting:
   a) a black king.
   b) an ace.
   c) a red card.
   d) a king or a queen.
   e) a black card.

Remark: Four groups (with two colors) 1 - 10 + King, queen, jack.

5. A bowel contains 17 red balls, 10 black, .10 white balls and 20 blue balls. If one of these is drawn at random, what the probabilities that it will be: a) a red b) a white c) a blue d) red or white e) white or blue f) neither white nor red.

6. Bowel I contains 12 red and 13 blue balls. Bowel II contains 10 red, 15 blue balls and 15 black balls. One ball is drawn from bowel I and placed in bowel II, then one ball is drawn from bowel II. Find probabilities that it will be:
   a) a red ball b) a blue ball
   c) a black ball.

5.5 Random Variable

In a typical population, it is usually possible to identify more than one characteristic of the units comprising it. For example, suppose the population is the collection of all full-time students registered at your university or college during the present academic term. In this instance, it is possible to identify numerous characteristics of the population unit-earned income, height, weight, sex, hair color, number of parking tickets accumulated during the term, grade-point average and so on. In a statistical study of the units in this population, we may be interested in just one characteristic (e.g…, gradepoint average) or in a collection of such characteristics (e.g…, sex and grade point average, or earned income and grade point average).

In Chapter 1, we referred to a population characteristic as a variable if the characteristic can assume one or more values in the population. We must now more specifically define a "variable" when we use the word to mean a measure of a population characteristic.

Definition 5.11
Random variable

A *random variable* is a numerically valued function whose value is determined by a random experiment.

A more mathematically rigorous definition is:

A *random variable* is a numerically valued function defined over a sample space.

Example 5.13

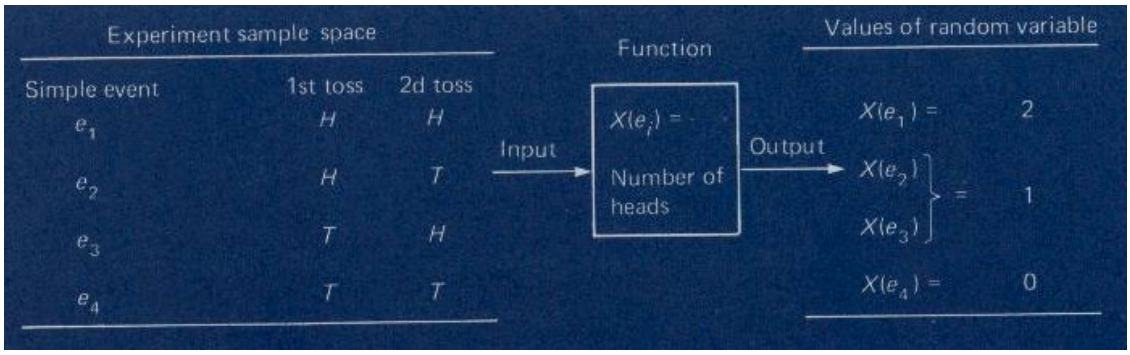
Suppose we consider the population of full-time students registered at your university or college. If we are interested in the population characteristic, "grade point average," and we select one student at random from this population, then the characteristic may be viewed as a random variable; its value is numeric and arises from a random experiment, and it is a function because it defines a correspondence between members of one set (the student population) and members of another set (the set of all possible grade point averages, from 0.00 to 4.00). For each student, the random variable defines one and only one grade point average, while more than one student

may have the same grade point average.

To better appreciate the concept of a random variable being a function. let us consider an example.

Example 5.14
Suppose the random experiment is tossing a coin twice, and we define the random variable, X = number of heads in the two tosses. figure 5.1 illustrates the correspondence between members of the experimental outcomes and possible values of the random variable.

Figure 5.7 Random variable X: number of heads in two tosses of a coin



Notice that the random variable X is a function. To each member in the first set (the simple events in the experimental sample space) there corresponds one and only one member in the second set (the values of the random variable). But each value of the random variable may correspond to one or more simple events. Notice that we have written the random variable in functional notation in figure 5.7 to emphasize its meaning.

Example 5.15
A production lot of 100 transistor radios contains 10 defectives. A retailer decides to select two of the radios at random, and by extensive testing, to determine whether they are defective. If neither radio is defective, he will accept the lot. Define the random variable, X = number of defective radios

(X = 0, 1, or 2). Determine the probabilities that X assumes each of its three possible values.

Solution:

In the experiment of selecting two radios at random, define the events:

A_1: First radio is defective

A_2: Second radio is defective

A_3: First radio is not defective

A_4: Second radio is not defective

The four simple events of the experiment are given in Table 5.1. The probability of each simple event occurring is determined by using the multiplicative law For example,

$$P(A_1 \cap A_2) = P(A_1)P(A_2/A_1) = (10/100)(9/99) = 0.0091$$

Notice that the two events are not independent; the outcome of the first selection affects the chance of the second radio being defective. Notice that the probabilities in Table 5.1 sum to one-we have specified the four mutually exclusive and collectively exhaustive simple events of the experiment.

Table 5.1 Outcomes and probabilities for Example 5.13

Simple events	Probability	Values of X
E_1:$A_1 \cap A_2$	**(10/100)(9/99)= 0.0091**	**2**
E_2:$A_1 \cap A_4$	**(10/100)(90/99)= 0.0909**	**1**
E_3:$A_3 \cap A_2$	**(90/100)(10/99)=0.0909**	**1**
E_4:$A_3 \cap A_4$	**(90/100)(89/99)= 0.8091**	**0**

Since the simple events are mutually exclusive and collectively exhaustive, P(X = 2) = 0.009 1, P(X = 1) = 0.0909 + 0.0909 = 0. 1818, and P(X = 0) = 0.8091. The values of the random variable X and their probabilities of occurrence are given in Table 5.2. This Table represents the probability distribution of the random variable X-a list of each value and its probability of occurrence.

Table 5.2
Probability distribution of X

X	P(x)
0	0.8091
1	0.1818
2	0.0091

From the probability distribution the probability that the retailer will accept the lot is 0.8091.

Table 5.3 Examples of discrete random variables

	Definition of X	*Values of X*	*Number of values of X*
1.	Number of correct answers by a student on a test containing 10 questions	0, 1, 2, . . . , 10	Finite (11)
2.	Number of tickets acquired by a student in a term. (Student is expelled after the 8^{th} ticket)	0, 1, 2, . . . , 8	Finite (9)
3.	Number of tickets acquired by a student in a term with no limit on the number received	0, 1, 2, . . .	Countably infinite
4.	Number of customers entering Ed's Hamburger Joint during one day's business	0, 1, 2, . . .	Countably infinite

Definition 5.12
Discrete random variable

A random variable is *discrete* if its set of values is finite or countably infinite in number.

A continuous random variable assumes values which occur on an interval or intersection of intervals on the real line. The number of values that a continuous random variable may assume is infinite. Examples of continuous random variables include height, weight, and the diameter of ball bearings. Although each of these variables is bounded (for example, the weight of individuals is bounded between zero lbs. and, say 500 lbs.), the variable can assume any of an infinite number of values between these bounds. Other examples of continuous random variables are given in Table 5.4.

Table 5.4 Examples of continuous random variables

Definition of X	Range of values of X
1. Diameter of 1/2 bolts produced in a machine shop	$X \geq 0$
2. Weight of a student in your class	$X \geq 0$
3. Amount of rainfall in inches recorded at a weather station on a given day	$X \geq 0$
4. Weight lost by a person weighing 300 lbs. on a diet designed for weight loss	$X \leq 300$ (A negative value of X indicates a weight gain.)

Definition 5.13

Continuous random variable

A random variable is *continuous* if it may assume all real number values in an interval.

5.6 Probablity Mass Function

The probability distribution can be described by a function P(x), call a probability mass function, which assigns probabilities to the values of discrete random variable.

Definition 5.14

Probability mass function (finite case)

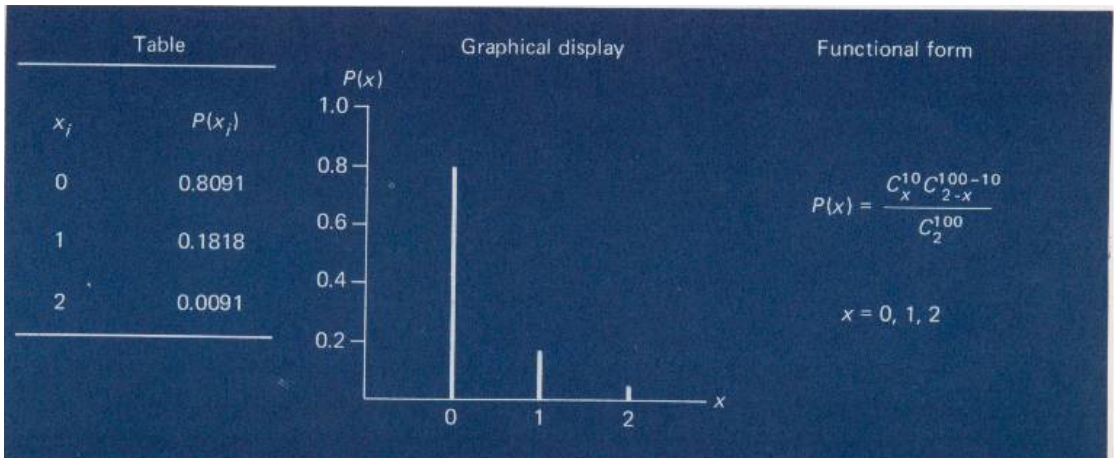
Let the random variable X assume a finite number of values, r in total, and denote these values by $x_1, x_2, ..., x_r$.

Let $P(x_i)$ be the probability that the random variable X assumes the value x_i. A *probability mass function* is a function which assigns probabilities to the values of a discrete random variable such that the following two conditions on the function P(x) are satisfied:

$$1.\ 0 \leq P(x_i) \leq 1\ ,\ i = 1,2, \ldots, r,$$

$$2.\ \sum_{i=1}^{r} P(x_i) = 1.$$

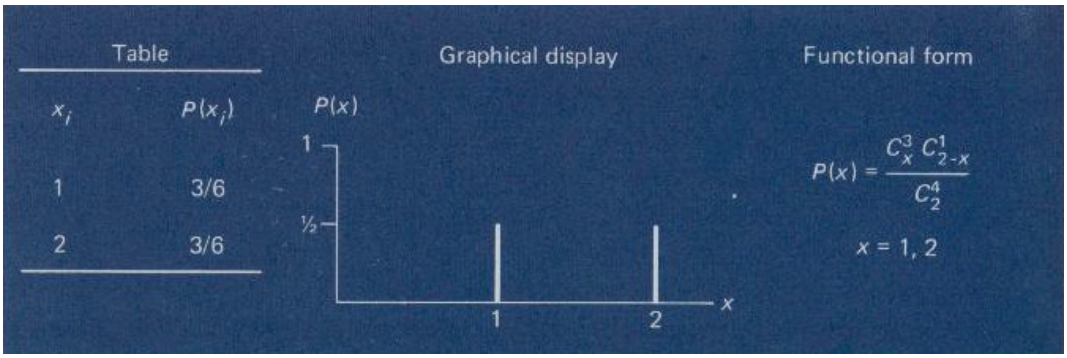
Figure 5.8 Three ways of presenting the distribution of the discrete random variable in Example 5.15



Example 5.16

A committee of two persons is to be formed from four persons, three of whom are female. The committee is formed anew at the end of each working week. The members of the committee have the duty of arriving at the office 30 minutes early to prepare the morning coffee, turn on lights, plug in machines and so on. In any given week, let X be the number of women serving on the present committee. Form the probability distribution of the random variable X.

Figure 5.9 Probability mass function for the random variable in Example 5.16



The average of the values of a random variable is called the expected value of the random variable and is given for a discrete random variable:

Definition 5.15

***Expected value* of a random variable (*discrete case*)**

Let X be a discrete random variable with a finite number of values denoted by $x_1, x_2; ..., x_n$. The mean or *expected value* of the random variable, denoted by E(X), is given by

$$E(X) = \sum_{i=1}^{r} x_i \cdot p(x_i) \ .$$

Example 5.17

Expected value of the random variable in Example 5.14 is given by the following table :

Table 5.7

xi	P(x)	xP(x)
0	0.8091	0.0000
1	0.1818	0.1818
2	0.0091	0.0182
Total		0.2000

Definition 5.16

***Variance* of a random variable (*discrete case*)**

Let X be a discrete random variable with a finite number of values denoted b) $x_1, x_2, ..., x_n$. The *variance* of the random variable, denoted by V(X), is given by

$$V(X)=\sum_{i=1}^{r} (x_i - E(x))^2 \ p(x_i) \ , \quad \text{or}$$

$$V(X) =\sum_{i=1}^{r} x_i^2 \ p(x_i) - [E(x)]^2 \ .$$

The first form of V(X) in definition 5.16 is a definitional form, and the second is the "computing form." Giving two expressions for V(X) is similar to giving two expressions for the population variance σ^2 in chapter 4.

Example 5.18

Compute the variance of the random variable in Example 5.15.

Solution:

The easiest way to compute the variance of a discrete random variable is by using a table similar to Table 5.8 below

Table 5.8 Partial computation of the variance of the random variable in Example 5.7

I	x_i	$P(x_i)$	x_i^2	$x_i^2P(x_i)$
1	0	0.8091	0	0.0000
2	1	0.1818	1	0.1818
3	2	0.0091	4	0.0364
				0.2182

From Table 5.7, E(X) = 0.2.

Thus

$$V(x)=\sum_{i=1}^{3} x_i^2P(x_i)- [E(x)]^2$$

$$= 0.2182 - (0.2)^2 = 0.1782.$$

The variance of X is O.1782 and the standard deviation of X, denoted by s is $(0.1782)^{1/2} = 0.422$.

5.7 Probability Density Function

Definition 5.17

Probability density function (continuous case)

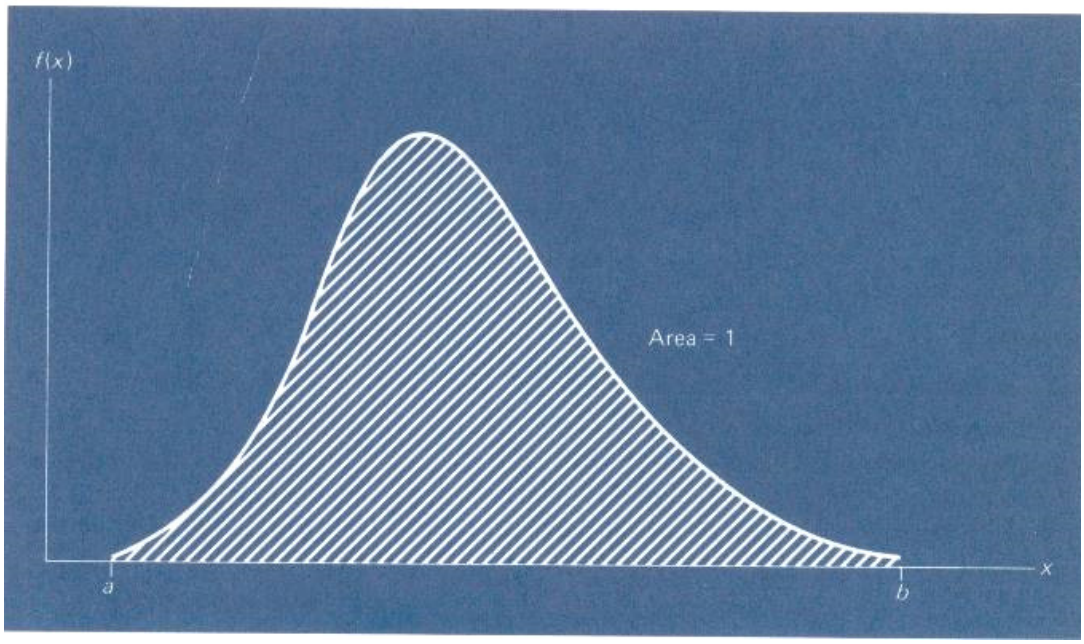
Let X be a continuous random variable defined over an interval of the real line from a to b, as illustrated in Figure 5.13.

The *probability density function* of X, denoted by f(x) , must satisfy two conditions:

1. $f(x) \geq 0$, $a \leq x \leq b$,

2. The area under f(x) from X= a to X= b must be one.

Figure 5.10 Probability density function f(x) of a continuous random variable

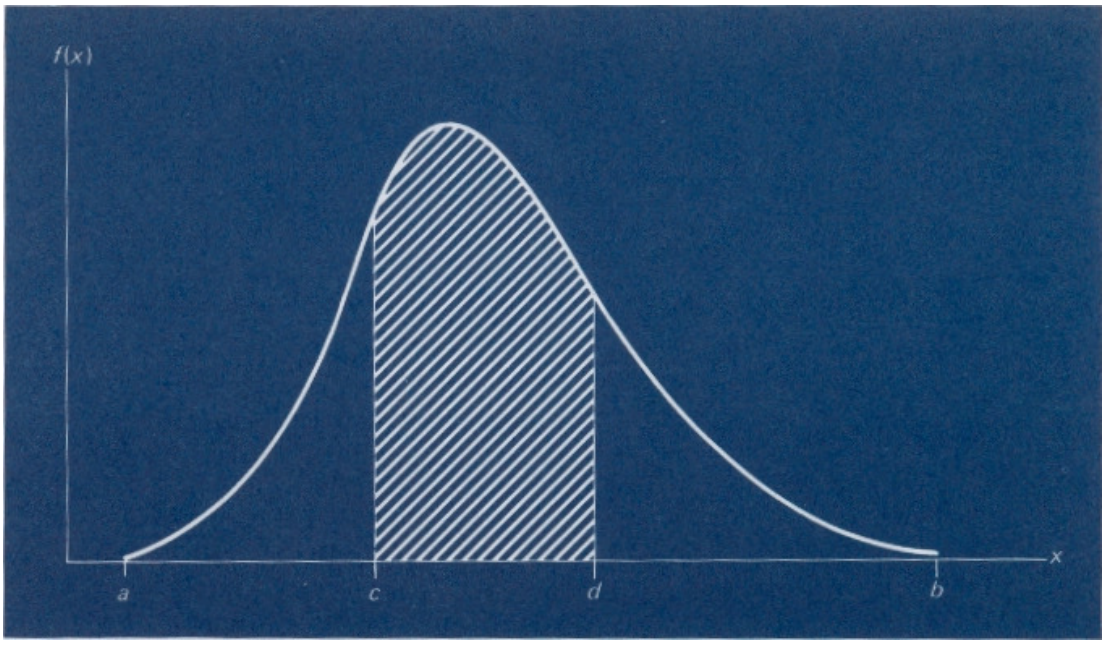


Notice the similarity between the conditions for P(x) to be a probability mass function of a discrete random variable given in definition 5.12 and the conditions placed on f(x) to be a probability density function for a continuous random variable given in definition 5.13. In the discrete case, probability is "massed" at the discrete values of the random variables while in the continuous case, the probability is spread "densely" over the range of the random variable. In the discrete case, the probability sticks must sum to one while in the continuous case, the dense set of sticks (the area under the function f(x) from a to b) must have an area of one unit.

Recall that if we desire to compute $P(c \leq X \leq d)$ where $d \geq c$ for a discrete random variable, we need only sum the probabilities of X taking on the values c through d; that is:

$$P(c \leq X \leq d) = \sum_{x=c}^{x=d} P(x) \quad , \quad d \geq c.$$

Figure 5.11 Area under the curve corresponding to $P(c \leq X \leq d)$



The analogy to a continuous random variable is straightforward. The probability that a continuous random variable X takes on a value between c and d is the area under f(x) between c and d as illustrated in Figure 5.11.

There is one significant difference in computing probabilities for a discrete and a continuous random variable. If the discrete random variable X can assume the value e, then the probability that X does assume this value is P(e), the height of the stick over e on the stick diagram for X. However, if e resides within the defined interval for a continuous random variable X, the probability that X assumes the value of e is zero; that is, $P(X = e) = 0$. This is true regardless of the numerical value of e over the defined interval of real numbers for the random variable X. Thus, $P(X = e)$ is not equal to f(e), the height of the curve at the point $X = e$.

The reason for this may be argued as follows. Let us assume that the continuous random variable is defined for values on the real line from $X = a$ to $X = b$. Assume that e is in the interval (a, b) and we let $P(X = e) = f(e)$. Since we have defined the probability that X assumes the value of e in (a, b) in this manner, this definition must also hold for all other values, say e_1, e_2,

e_3, . . . , in the interval (a, b). Since there is an infinite number of values (not countable) between (a, b), we can see that the sum of the "probabilities" f(e) + f(e_1)+ f(e_2) + f(e_3) +••• will quickly exceed one, which means that any particular number, say f(e_2), can no longer be interpreted as a probability. Indeed, a single f(e_i) may exceed one by itself if the height of the curve at the point X = e; is greater than one.

Another way to look at this is to write P(X = e) as P(e ≤ X≤ e) and use the definition for the probability that X assumes a value between two points, say c and d, as illustrated in figure 5.9. Since there is no area between e and e, P(X = e) = P(e ≤ X≤ e) = 0.

Table 5.9 Comparison of properties of discrete and continuous random variables

Property	*Discrete random variable*	*Continuous random variable*
	Finite *or* countably infinite	Infinite (not countable)
Prob($c \le X \le d$)	$\sum_{x=c}^{x=d} P(x)$	Area under the curve $f(x)$ from c to d $= \int_c^d f(x)\,dx$
Prob($X = e$)	$P(e)$	*Always Zero* $\text{Prob}(X = e) = \int_e^e f(x)\,dx \equiv 0$

Example 5.18
Consider the probability density function

$$f(x)=\begin{cases} 1 & 0 \le x \le 1\ , \\ 0 & x<0 \text{ or } x>1. \end{cases}$$

Find :

1. P(0 ≤ x ≦ 1)
2. P(0.25 ≤ x ≤ 0.60)
3. P(x > 0.75)

Solution:

1. $P(0 \leq x \leqq 1) = \int_0^1 f(x)\,dx$

   $= \int_0^1 1\,dx = 1.$

2. $P(0.25 \leq x \leq 0.60) = \int_{0.25}^{0.6} f(x)\,dx$

3. $P(x > 0.75) = \int_{0.75}^{1} f(x)\,dx$

Definition 5.18

***Expectation* of a random variable (*continuous case*)**

Let X be a continuous random variable defined over the interval (a, b) with probability density function f(x). The mean or *expected value* of the random variable, denoted by E(X), is given by

$$E(X) = \int_a^b x\, f(x)\,dx.$$

Definition 5.19

***Variance* of a random variable (*continuous case*)**

Let X be a continuous random variable defined over the interval (a, b) with probability density function f(x). The *variance* of the random variable, denoted by V(X), is given by

$$V(X) = \int_a^b [x - E(x)]^2 f(x)dx, \text{ or}$$

$$V(X) = \int_a^b x^2 f(x)\,dx - [E(x)]^2 .$$

It is interesting to note the similarity in the computation of the expected value and the variance of a discrete and a continuous random variable as illustrated in Table 5.10. The analogy to summing in the discrete case is integrating in the continuous case .

Table 5.10 Comparison of the expected value and variance formulas for discrete and continuous random variables

Property	Discrete random variable	Continuous random variable
Expected Value		
$E(X)$	$\Sigma\,(xP(x))$	$\int x\,f(x)\,dx$
Variance		
$V(X)$	$\Sigma\,(x - E(X))^2P(x)$	$\int (x - E(X))^2 f(x)\,dx$

Example 5.19

Let

$$f(x) = \begin{cases} 0.5, & 0 \le x \le 2, \\ 0, & \text{otherwise.} \end{cases}$$

Find:
a. E(x)
b. Var (x)

Solution :

$E(x) = \int_0^2 1/2\ x\ dx = x^2/4\ ]_0^2 = 1$,
$Var(x) = \int_0^2 1/2\ x^2\ dx - (1)^2 = 2/6.$

Tutorial 5.2

I . In a certain population of voters, it is known that 60 percent are Democrats and 40 percent are Republicans. If a sample of' three voters is extracted from this population, find the probability distribution of the random variable, X = Number of Democrats in the sample.

2. An automobile salesperson has a probability of 0.20 of selling a car to each individual interested in buying a car with whom she speaks on the showroom floor. On a certain day, the salesperson talks with four individuals regarding the purchase of a car. Find the probability distribution of the random variable, X = Number of cars sold. Show the distribution in the form of a table, a stick diagram, and (if possible) a formula.

3. A multiple choice exam consists of four questions, each of which has four possible answers. If a student is forced to guess on all four questions, what is the probability distribution of the random variable, X = Number of correct guesses?

4. In an organization consisting of 5 women and 10 men, a committee of four individuals is to be selected at random from the 15 people. Find the distribution of the random variable, X = Number of women on the committee.

5. In a six-cylinder automobile engine. two spark plugs are defective. Three spark plugs are removed at random and checked. Let X be the number of defectives found (0, 1, or 2). Find the probability distribution of X.

6. Suppose two dice are tossed. Let X be the sum of the dots on the top faces of the two dice. Find the probability distribution of X.

7. Let X be a random variable with probability distribution given by the following table:

X	P(x)
0	0.70
1	0.20
2	0.06
3	0.04

Find:
a. The expected value of X.
b. The variance of X.
c. The mode of X.
d. The range of X.

8.It is found that the probability distribution of X is:

X	P(x)
0	0.95
1	0.03
2	0.015
3	0.003
4	0.0015
5	0.0005

Find:

a. The expected value of X.
b. The variance of X.
c. The mode of X.
d. The range of X.

9. Consider the formula: $P(x) = x^2/21$, x = 0, 1, 2, 4. Is P(x) a probability mass function? If so, show the distribution of X in tabular form and compute the expected value of X.

10. Simulate the experiment in Problem 5 by taking six scraps of paper and marking on two of them, the letter "D." Place the six scraps of paper in a box and randomly select three of the scraps. Record the number marked D.Do this experiment 100 times. Construct a frequency distribution for this sample and compare it with the theoretical probability distribution determined in Problem 5.

11. Compute the expected value of the random variable X in Problem 5. Find the sample mean number of defectives recorded per trial in the

Problem 10 simulation experiment. Does X, the sample mean, provide a good estimate of E(X)" Should X provide a good estimate of E(X)? Why?

12. From the most recent national census, it is found that the number of children (X) in American families follows the following probability distribution:

Number of children, X	**Proportion of families. P(x)**
0	**0.48**
1	**0.20**
2	**0.15**
3	**0.08**
4	**0.05**
5	**0.03**
6	**0.01**

It is assumed that the proportion of families with more than six children is negligible.

a. Find the expected value and standard deviation of X.

b. Form a stick diagram of this distribution. Is the distribution skewed?

c. find var(X).

13. Consider the following functions

$$f(x)=\begin{cases}(2/3)\,x & -1< x< 2\\ 0 & \text{otherwise}\end{cases},$$

$$f(x)=\begin{cases}1/5, & -1 < x < 4,\\ 0, & \text{otherwise.}\end{cases}$$

Calculate :

a.p(-1<x<0) ,

b.p(X>1) ,

c.p(X=2),

d.E(X) ,

e.var(X) .

6

Binomial & Normal distributions

6.1 Probability Function

The following tables, serve to illustrate what we mean by a probability function, namely, a correspondence which assigns probabilities to the values of a random variable.

Example 6.1

The first of the two tables which follow was easily obtained on the basis of the assumption that each face of the die in question has a probability of 1/6 and the second was obtained by considering as equally likely the eight possible outcomes HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT of three flips of a coin, where H stands for heads and T for tails:

Table 6.1

Number of Points Rolled With a Die	*Probability*
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Table 6.2

Number of Heads Obtained in Three Flips of a Coin	*Probability*
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

Note that in each of these examples the sum of all the probabilities is 1.
Note also that since the values of probability functions are probabilities, they must always be positive or zero, and cannot exceed 1.

Whenever possible, we try to express probability functions by means of formulas which enable us to calculate the probabilities associated with the various values of a random variable. With the usual functional notation.

we can thus write

$$f(x)=1/6 \qquad \text{for } x=1,2,...,6,$$

for the first of the above examples, where f(1) represents the probability of rolling a 1, f(2) represents the probability of rolling a 2, and so on.

6.2 The Binomial Distribution

There are many applied problems in which we are interested in the probability that an event will take place x times in n "trials," or in other words, x times out of n, while the probability that it will take place in any one trial is some fixed number p and the trials are independent. We may thus be interested in the probability of getting 24 responses to 80 mail questionnaires, the probability that in a sample of 50 voters 32 will favor Candidate A, the probability that 3 of 10 laboratory mice react positively to a new drug, and so on. Referring to the occurrence of any one of the individual events as a "success", we are thus interested in the probability of getting x successes in n trials. To handle problems of this kind, which incidentally include, we use a special probability function, that of the binomial distribution.

If p denotes the probability of a success on any given trial, the probability of getting x successes in n trials (and hence, x successes and n - x failures) in some specific order is $p^x (1 - p)^{n-x}$. There is one factor p for each success, one factor 1 - p for each failure, and the x factors p and n - x factors 1 - p are all multiplied together by virtue of the assumption that the n trials are independent. Since this probability is the same for each point of the sample space where there are x successes and n- x failures (it does not depend on the order in which the successes and failures are obtained), the desired probability for x successes in n trials in any order is obtained by multiplying $p^x (1-p)^{n-x}$ by the number of points of the sample space (that is, individual outcomes) where there are x successes and n - x failures. In other words, $p^x (1-p)^{n-x}$ is multiplied by the number of ways in which the x successes can be distributed among the n trials, namely, by n!/x!(n-x)! we have thus arrived at the following result:

Definition 6.1
Binomial distribution

The probability of getting x successes in n independent trials is given by

$$f(x)=\binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } n=0,1,2,\ldots,n,$$

where p is the constant probability of a success for each individual trial.

It is customary to say that the number of successes in n trials is a random variable having the binomial probability distribution, or simply the binomial distribution. The terms "probability distribution" and "probability function" are often used interchangeably, although some persons make the distinction that the term "probability distribution" refers to all the probabilities associated with a random variable, and not only those given directly by its probability function. Incidentally, we refer to this distribution as the binomial distribution because for x = 0, 1, 2, . . ., and n, the values of its probability function are given by the successive terms of the binomial expansion of $((1 - p) + p)^n$.

Example 6.2

To illustrate the use of the above formula, let us first calculate the probability of getting 5 heads and 7 tails in 12 flips of a balanced coin. Substituting x = 5, n = 12, p=1/2, and (12!/5!.7!)=792

$$f(5)= 792\ (1/2)^5\ (1-1/2)^{12-5}= 99/512,$$

or approximately 0.19. Similarly, to find the probability that 7 of 10 mice used in an experiment will react positively to a drug, when the probability that any one of them will react positively is 4/5 we substitute x = 7, n = 10, p =4/5 and (10!/7! 3!)= 120 , and we get

$$f(7)=120(4/5)^7(1-4/5)^{10-7}$$

or approximately 0.20 .

Remark 6.1
Some of probability values of binomial distribution are given in Table 1(in appendix).

Example 6.3

To give an example in which we calculate all of the values of a binomial distribution, suppose that a safety engineer claims that only 60 per cent of all drivers whose cars are equipped with seat belts use them on short trips. Assuming that this figure is correct, what are the probabilities that under such conditions 0, 1, 2, 3, 4, or 5 of 5 drivers will be using their seat belts? Substituting n = 5, p = 0.60, and, respectively, x= 0, 1, 2, 3, 4, and 5, we use

$$f(x)=\binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } n=0,1,2,3,4,5,$$

Then

$$f(0) = \binom{5}{0}(0.60)^0(1-0.60)^{5-0} = 0.010$$
$$f(1) = \binom{5}{1}(0.60)^1(1-0.60)^{5-1} = 0.077$$
$$f(2) = \binom{5}{2}(0.60)^2(1-0.60)^{5-2} = 0.230$$
$$f(3) = \binom{5}{3}(0.60)^3(1-0.60)^{5-3} = 0.346$$
$$f(4) = \binom{5}{4}(0.60)^4(1-0.60)^{5-4} = 0.259$$
$$f(5) = \binom{5}{5}(0.60)^5(1-0.60)^{5-5} = 0.078$$

where all the answers are rounded to three decimals. A histogram of this distribution is shown in figure 6.1.

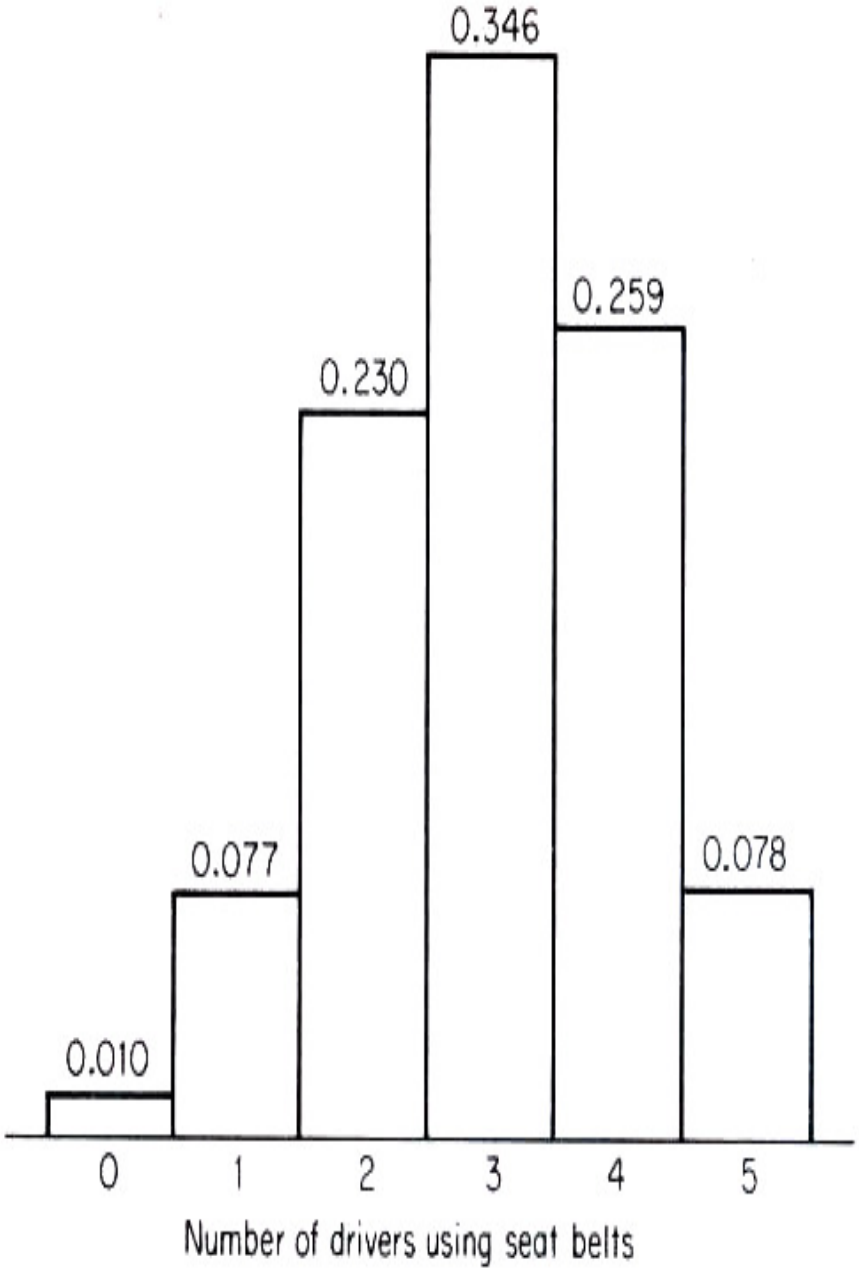
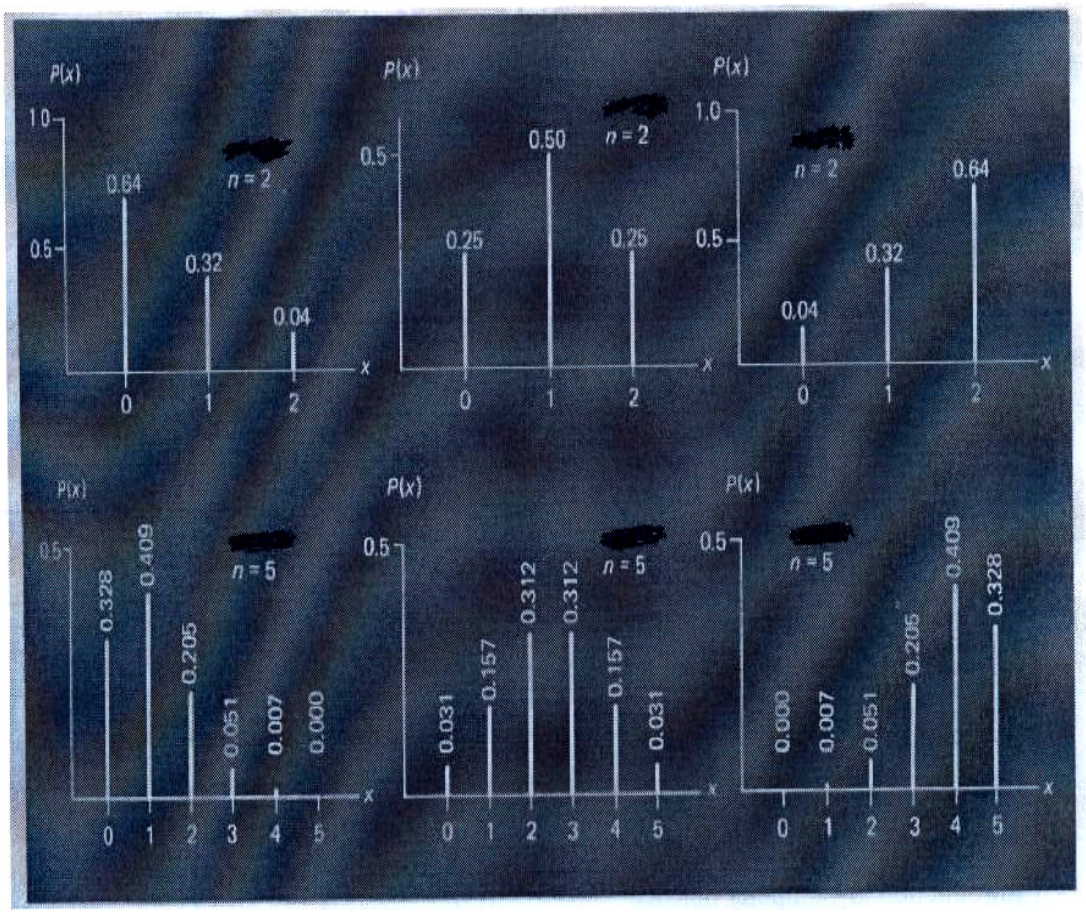


Figure 6.1 Histogram of binomial distribution with n=5 and p=0.60.

Figure 6.2 Some specific members of the binomial distributions where :
a- p = 0.2 , b- p = 0.5 , c- p = 0.8 ,



d- p = 0.2 , e- p = 0.5 , f- p = 0.8 .

<u>Remark 6.2</u> If X is a binomial random variables then its mean E(x) and variance , V(x) are given by:

$$E(x) = n.p \ , \ var(x) = n.p(1-p)$$

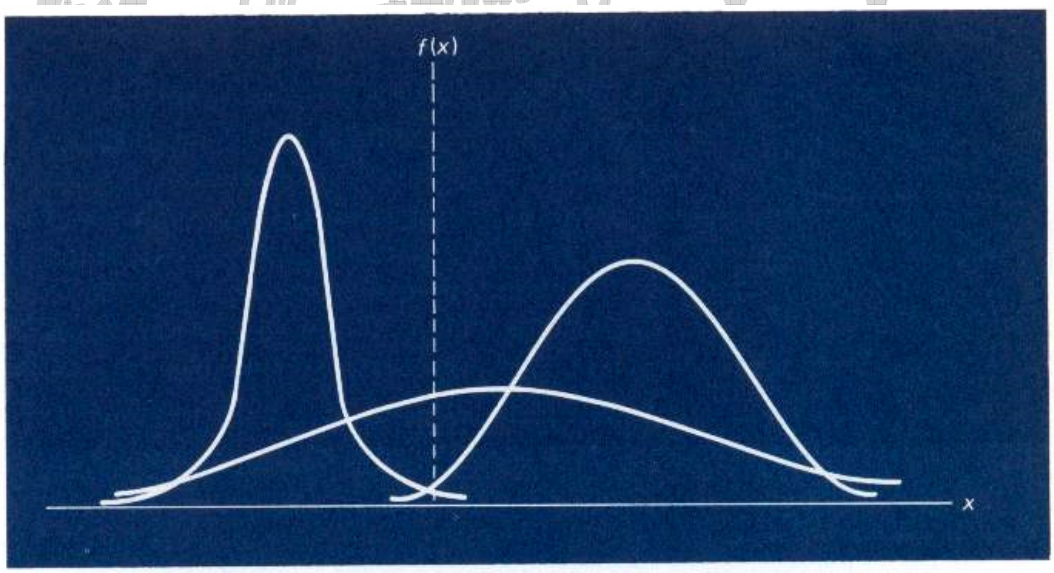
Example 6.4
Consider figure 6.2 find the mean and the variance for each p.m.f.
Solution:
i) E(x)=2(0.2) ,
Var(x)=2 (0.2)(0.8) ,

ii) E(x)=2(0.5) ,
Var(x)=2(0.5)(0.5) ,
iii) E(x)=2(0.8) ,
Var(x)=2(0.8)(0.2),
iv) E(x)=5(0.2) ,
Var(x)=5(0.2)(0.8) ,
v) E(x)=5(0.5) ,
Var(x)=5(0.5)(0.5) ,
vi) E(x)=5(0.8),
Var(x)=5(0.8)(0.2) .

6.3 The Normal Distribution

The normal distribution is "probably" the most important probability distribution in statistics. It is a probability distribution of a continuous

Figure 6.3 Three forms of the normal distribution



distribution is that of a bell-it has a single mode and is symmetric about its central value. The flexibility in using the normal distribution is due to the fact that the curve may be centered over any number on the real line and that it may be made flat or peaked to correspond to the amount of dispersion in the values of a random variable. Many quantitative characteristics have distributions similar in form to the normal distribution's bell shape.

Examples of random variables that have been modeled successfully by the normal distribution are the height and the weight of people, the diameters of bolts produced by a machine, the IQ of people, the life of batteries or light bulbs, and so on. Typically, the type of experiment that produces a random variable that can be successfully approximated by a normal random variable is one in which the values of the random variable are produced by a measuring process, where it is known that the measurements tend to cluster symmetrically about a central value. A random variable that is an average or a sum of values of another random variable is, under very general conditions, almost always distributed approximately as a normal random variable, regardless of the form of the distribution of the random variable whose values are summed or averaged. An example of such a random variable is the mean grade point average of a randomly selected group of students. The notion that a random variable that is an average is distributed as a normal random variable is discussed in the next chapter with the central limit theorem.

Unfortunately, if it is known that the distribution of a random variable is symmetrically distributed with a single mode, the random variable may not necessarily be a normal random variable. There are other distributions in statistics that are unimodal and symmetric. However they also can often be modeled successfully by the normal distribution. For a random variable to be normally distributed, the mathematical expression delineating the form of the bell must be of a specific type, as described in the following definition.

Definition 6.2

Normal random variable and its probability density function

A continuous random variable X is said to be normally distributed if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]} \quad \text{where} \quad -\infty < x < +\infty, \quad -\infty < \mu < +\infty, \quad \sigma > 0$$

where μ and σ are parameters of the distribution and π and e are mathematical constants equal to 3.14159 and 2.71828, respectively.

6.3.1 Mean and variance of the normal random variable

The mean and variance of the normal random variable may be determined by performing the following integrations:

$$\text{Mean:} \quad E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}\,dx$$

$$\text{Variance:} \quad V(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)\,dx$$

$$= \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}\,dx$$

As might be suspected, these integrals are not easily evaluated. The results of integration are rather simple, however, and are given in the following theorem.

Remark 6.3

Mean and variance of the normal random variable

If X is a normal random variable, then its *mean*, $E(X)$, and *variance*, $V(X)$, are given by:

$$E(X) = \mu$$

$$V(X) = \sigma^2$$

Notice that the mean depends only on the parameter μ, and that the variance depends only on the parameter σ. Thus, the normal distribution may be located over its central value on the real line independently of the amount of dispersion σ^2 specified for the distribution. Contrast this with the binomial distribution (and others discussed thus far) in which the mean and the variance both depend upon the parameters n and p, [E(X) = n.p , V(X) = n.p.(1-p)] and hence are not independent of one another. This property of the normal distribution adds immeasurably to its flexibility in modeling the distributions of non-normal random variables.

We will now return to the problem of computing probabilities associated with a normal random variable.

6.3.2 Standardized normal distribution

Probabilities associated with any member of the normal distribution family can be computed from a table of probabilities compiled for the standard normal distribution.

Definition 6.3

Standard normal distribution

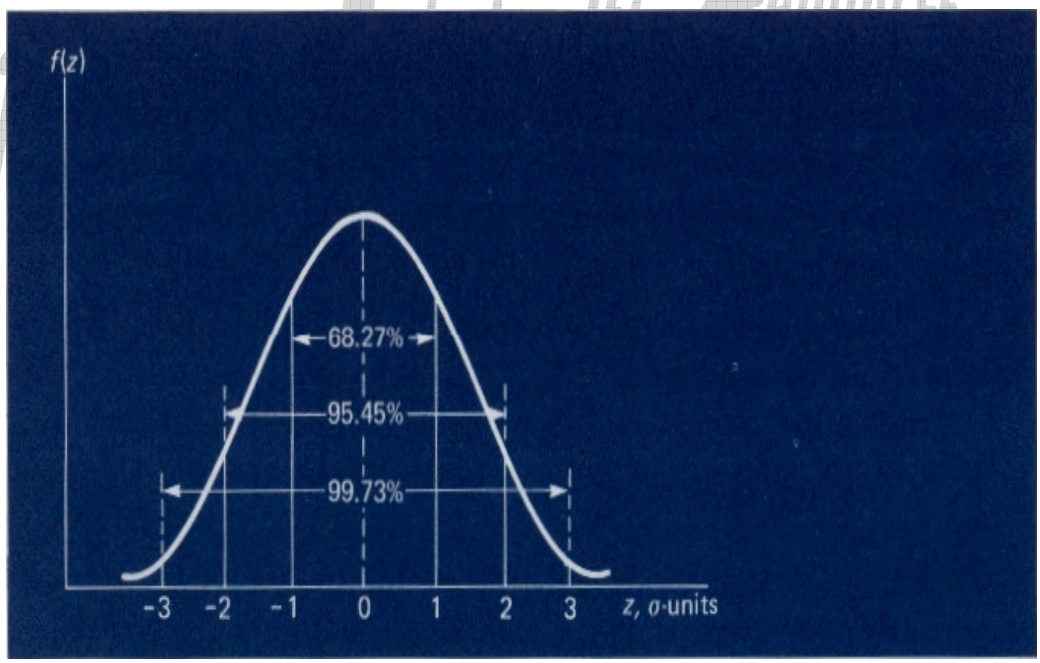
A normal distribution with μ= 0 and σ = 1 is called a standard normal distribution. When a normal random variable X has a mean of zero and a variance of one, it will be called a standardized normal random variable and will be denoted by Z. The probability density function of the standardized normal random variable Z is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{1}{2}(z^2)\right]}, \text{ where } -\infty < z < \infty$$

The form of the standard normal distribution is illustrated in figure 6.4. As indicated in figure 6.4, and for any normal distribution, 68.27 percent of the values of z lie within one standard deviation of the mean, 95.45 percent of the values lie within two standard deviations of the mean, and 99.73 percent of the values lie within three standard deviations of the mean.

Probabilities of a standardized normal random variable of the form P ($0 \leq Z \leq a$) are provided in Table 2(appendix) . By using the fact that the normal distribution is symmetric about its mean (zero in this case), and that the total area under the curve is one (half to the left of zero, and half to the right), the probability that Z resides in any interval on the real line may be determined from this table, as the following example indicates.

Figure 6.4 Standard normal distribution



Example 6.5

Find the area under the standard normal distribution curve for each of the intervals listed below.

a. Between Z = 0 and Z = 2.0

b. Between Z=-1.28 and Z = 0.0

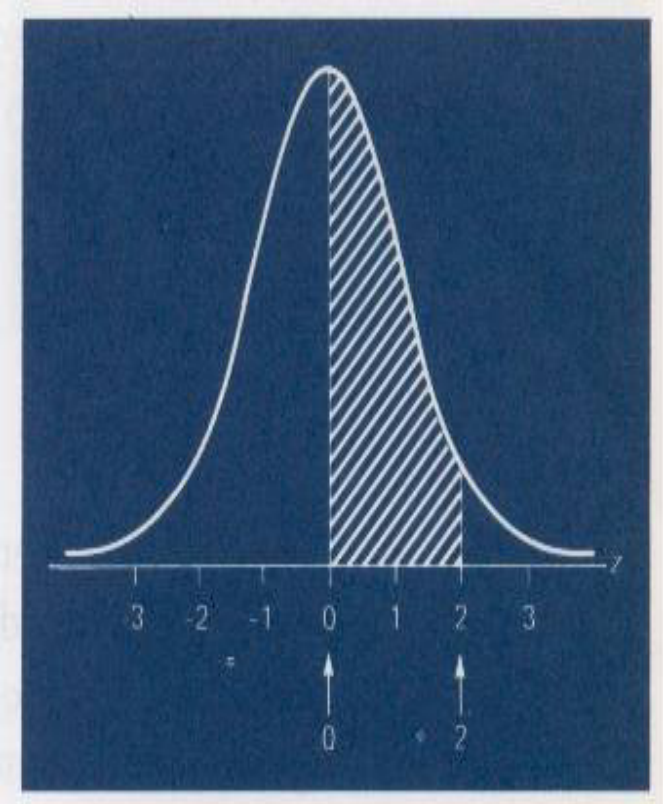
c. Between Z=-0.58 and Z= 2.54

d. Between Z = 1.20 and Z = 2.4

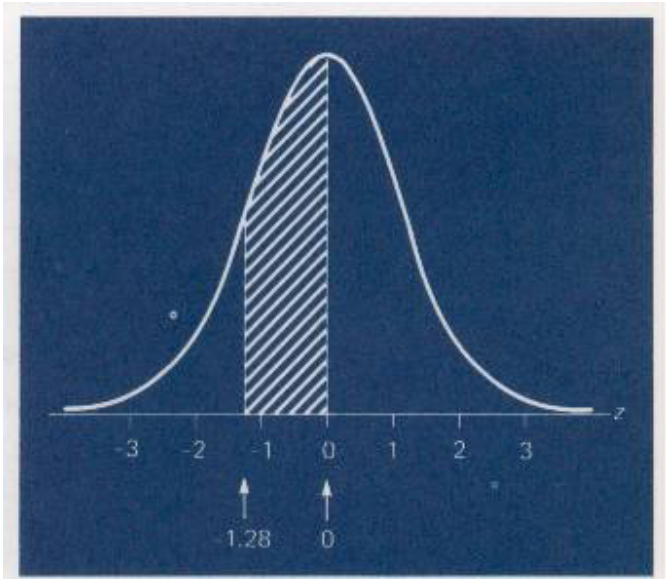
e. Greater than Z = 2.87

Solution:

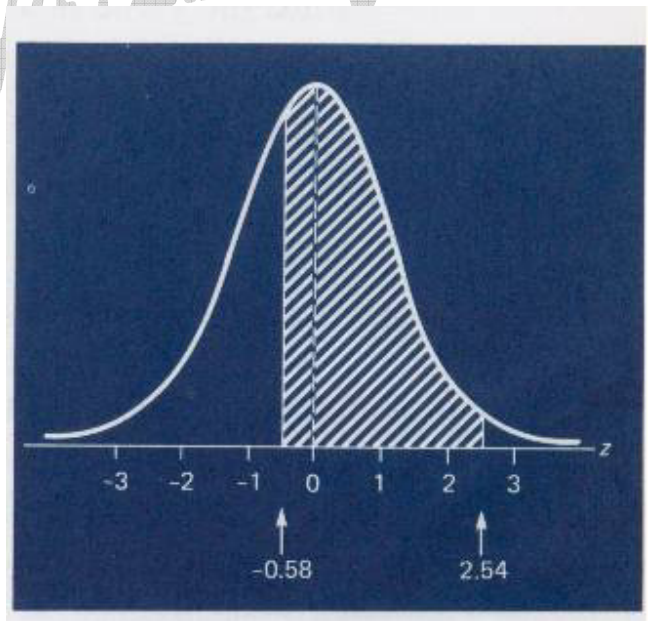
a. In Table 1, proceed downward in the leftmost column until 2.0 is reached. Select the first column marked .00 indicating that the second decimal place is zero. The area read from the table is 0.4772.



b. Since the normal distribution is symmetric, the area between - 1.28 and 0.0 is equal to the area between 0.0 and 1.28. Thus, proceed down the leftmost column until 1.2 is reached. Select the ninth column marked 0.08. The resulting number in the table is 0.3997.

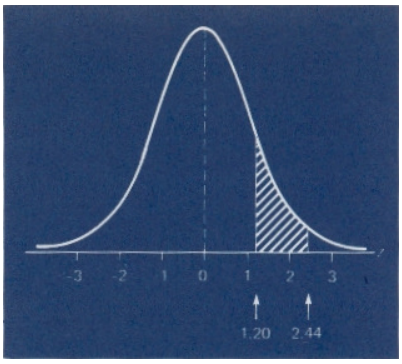


c. We may determine this area in two parts: total area = (area between 0.0 and 2.54) + (area between -0.58 and 0.0). The area between 0.0 and 2.54 is 0.4945 from Table 1. The area between -0.58 and 0.0 is the same as the area between 0.0 and 0.58, which is 0.2190. The answer is 0.4945 + 0.2190 = 0.7135.

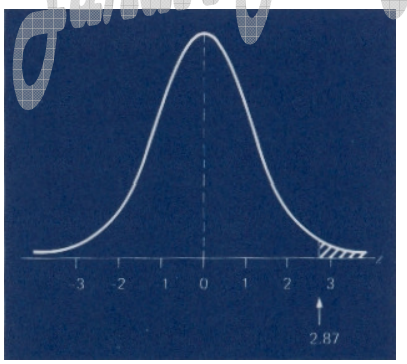


d. We may determine this area by differencing the area from 0 to 2.44 and from 0 to 1.20. The area between 0 and 2.44 is 0.4927, and the area

between 0 and 1.20 is 0.3849. Thus, the area between 1.20 and 2.44 is 0.4927 - 0.3849 = 0.1078.



e. Since the area between 0 and + ∞ is 0.5, we can determine the area from 2.87 to ∞ by subtracting the area from 0 to 2.87 (0.4979) from 0.5: 0.5000 -0.4979=0.0021.



In many problems, we will be given the area in a certain interval and be asked to determine the value of Z that specifies the interval. This is the reverse of the problems solved in Example 6.5 demonstrates the use of Table 1 to solve the "reverse" problem.

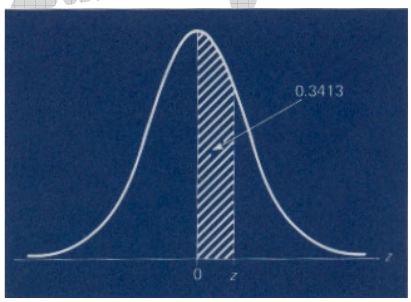
Example6.6

Find the value of Z on the standard normal distribution axis for each of the areas listed below.

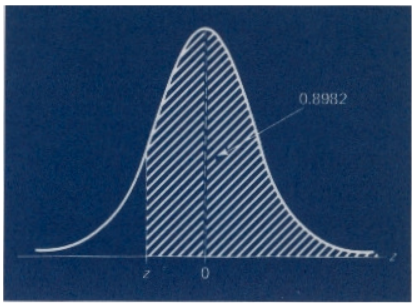
a. The area between 0 and z is 0.3413 b. The area to the right of z is 0.8982

Solution:

a. Table 1 must be used in reverse. We look in the body of the table for the area 0.3413. It appears in the row marked 1.0 and the column marked 0.0. Thus, the value of Z is 1.00.



b. Since the area given is greater than 0.5, we know that z must be less than zero. The area between z and 0 is 0.8982 - 0.5000 = 0.3982. Now assume that z is positive and find z so that the area between 0 and z is 0.3982. The area of 0.3982 does not appear in the tables; the closest numbers are 0.3980 and 0.3997. The exact value of z could be determined by interpolation, but we will use z= 1.27 since 0.3980 is closer to 0.3982 than is 0.3997. We must remember that z must be to the left of zero (a negative number). Thus, z = -1.27.



6.3.3 Areas under the normal distribution

Probabilities associated with a normal random variable X that is not standardized can be determined from Table 1 by using the results of the following theorem.

Theorem 6.1
Standardization of a normal random variable

If X is a normal random variable, the mean of which is μ and the standard deviation of which is σ, then

$$Z=(x-\mu)/\sigma ,$$

is a standardized normal random variable with a mean of zero and a standard deviation of one.

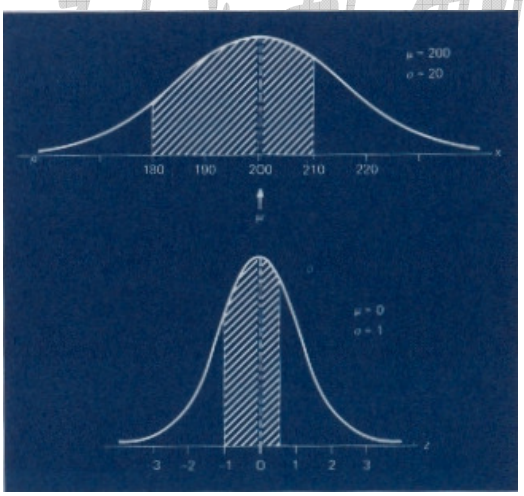
The following examples illustrate the use of Theorem 6.1, and more generally, the applicability of the normal distribution model.

Example 6.7

The mean lifetime of 50-watt lightbulbs produced by the Stay-Bright Lightbulb Company is 200 hours. It is known that the standard deviation is 20 hours. Assuming that the lifetimes of the lightbulbs are normally, distributed, what are the probabilities that a single 50-watt lightbulb extracted from the production lot will

a. Burn out between 180 hours and 210 hours?

b. Burn out at a time greater than 250 hours?



Solution:

a. The solution on the X distribution with a mean of 200 and a standard deviation of 20 is the area between X = x_1 = 180 and X=x_2= 210. This

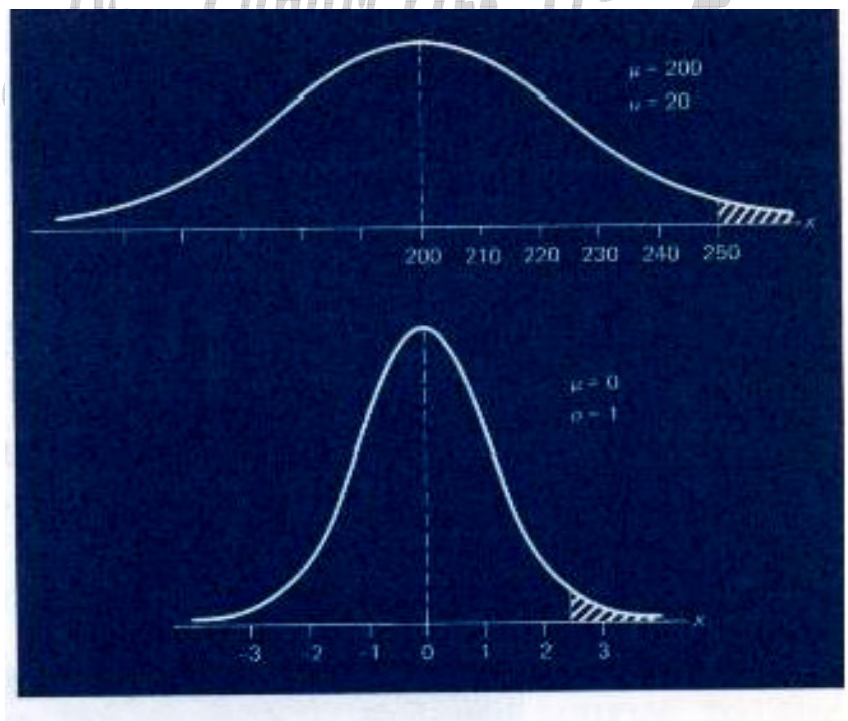
area can he determined by first standardizing x_1 and x_2:

$$Z_1 = (x_1 - \mu)/\sigma = (180 - 200)/20 = -20/20 = -1.00$$

The area between - 1.0 and 0.50 on the standard normal distribution will equal the area between 180 and 210 on the X distribution. The area between $z_1 = -1.00$ and $z_2 = 0.50$ from Table 1 is:

Area(- 1.00 to 0.00) = 0.3413
+ Area(0.00 to 0.50) = 0.1915
0.5328

Thus , Prob (180 $\leq$ X $\leq$ 210) = P(-1.00$\leq$Z$\leq$0.50)= 0.5328. This answer tells us that 53.28 percent of the 50-watt lightbulbs comprising the



population will fail between 180 and 210 hours.

b. The solution on the X distribution is the area greater than $x_1 = 250$ hours. The standardized value is

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{250 - 200}{20} = \frac{50}{20}$$

$$= 2.5$$

The area from $z_1 = 2.5$ to $+\infty$ is:

$$0.5000 - \text{Area}(0.00 \text{ to } 2.5)$$
$$= 0.5000 - 0.4938 = 0.0062$$

Thus, 0.62 percent of the 50-watt lightbulbs will fail at a time exceeding 250 hours.

Tutorial 6

1. Using the normal probability tables, calculate the areas under the standard normal curve for the following z values:
   a. Between $Z = 0.0$ and $Z = 1.2$
   b. Between $Z = 0.0$ and $Z = -0.9$
   c. Between $Z = 0.0$ and $Z = 1.45$
   d. Between $Z = 0.0$ and $Z = -1.44$
   e. Between $Z = 0.3$ and $Z = 1.56$
   f. Between $Z = -1.71$ and $Z = -2.03$
   g. Between $Z = -1.72$ and $Z = 2.53$
   h. Between $Z = -0.02$ and $Z = 3.54$
   i. Greater than $Z = 2.50$
   j. Greater than $Z = -0.60$
   k. Less than $Z = -1.22$
   l. Less than $Z = 1.66$

2. Find the value of z_0 on the standard normal curve so that
   a. $\text{Prob}(Z \geq z_0) = 0.60$
   b. $\text{Prob}(Z \leq z_0) = 0.02$
   c. $\text{Prob}(Z \geq z_0) = 0.20$
   d. $\text{Prob}(Z \leq z_0) = 0.85$
   e. $\text{Prob}(-z_0 \leq Z \leq z_0) = 0.95$
   f. $\text{Prob}(-z_0 \leq Z \leq z_0) = 0.20$

3. Check whether the following can be looked upon as probability functions (defined in each case only for the given values of x) and explain your answers:

   (a) $f(x) = 1/4$ for x= 0, 1, 2, 3, or

   (b) $f(x) = (x+1)/10$ for x = 0, 1, 2, or 3;

   (c) $f(x) = (x-2)/5$ for x = 1, 2, 3, 4, or 5;

   (d) $f(x) = x^2/30$ for x = 0, 1, 2, 3, or 4.

4. In each case check whether the given values can be looked upon as the values of the probability function of a random variable which can take on only the values 1, 2, 3, and 4, and explain your answers:

(a) f(1) = 0.24, f(2) = 0.24, f(3) = 0.24 , f(4) = 0.24;

(b) f(1) = 1/6, f(2) = 2/6, f(3) = 2/6 , f(4) = 1/6;

(c) f(1) = 0. 13, f(2) = 0.38, f(3) = 0.04 , f(4) = 0.45;

(d) f(l) = 1/2, f(2) = 1/4, f(3) = 1/8 , f(4) = 1/16.

5. Use the formula for the binomial distribution to find the probability of getting :

(a) exactly 3 heads in 8 flips of a balanced coin; (b) at most 3 heads in 8 flips of a balanced coin; (c) exactly 1 one in 3 rolls of a balanced die; (d) at most 1 one in 3 rolls of a balanced die,(e) calculate E(x) and Var(x) .

6. A multiple-choice test consists of 8 questions and 3 answers to each question (of which only one is correct). If a student answers each question by rolling a balanced die and checking the first answer if he gets a 1 or a 2, the second answer if he gets a 3 or a 4, and the third answer if he gets a 5 or a 6, find (by means of the formula for the binomial distribution) the probability of getting

(a) exactly 3 correct answers;
(b) no correct answers;
(c) at least 6 correct answers.

7

Regression Analysis

7.1 Introduction

Linear regression analysis is a technique used to predict the value of one quantitative variable by using its relationship with one or more additional quantitative variables. For example, if we know the relationship between height and weight in adult. males, we can use regression analysis to predict weight given a particular value for height.

The relationship between height and weight is familiar to us; generally, the taller a person is, the more he weighs. Another example of a familiar relationship is that of crop yield and the amount of fertilizer applied to the land; the more fertilizer applied to the land, the greater the yield-to a point. If too much fertilizer is applied, the crop will be killed off by the fertilizer chemicals-the land will be "burned." An important relationship in business is the relationship between the allocation of dollars to advertising effort and the level of sales of a product; the more money expended in advertising, the greater the level of sales (in general).

In this chapter, we will emphasize the development of regression analysis when a single predictor variable is used to predict the variable of interest, and where the relationship between the variables is linear. In this context, the variable which is used to predict the variable of interest is called the *independent variable,* and the variable we are trying to predict is called the *dependent variable.* The analysis used is called ***simple linear regression analysis-****simple* because there is only one predictor or independent variable, and *linear because* of the assumed linear relationship between the dependent and the independent variables.

Certainly, it is common to find that the variable of interest is related to more than one predictor or independent variable, or that the relationship is not linear. An example is the level of sales of a product and the advertising expenditure. The sales level of a product generally "depends" upon more predictor variables than advertising expenditure alone. However, we will often find that quite good predictions are possible based upon a single predictor variable.

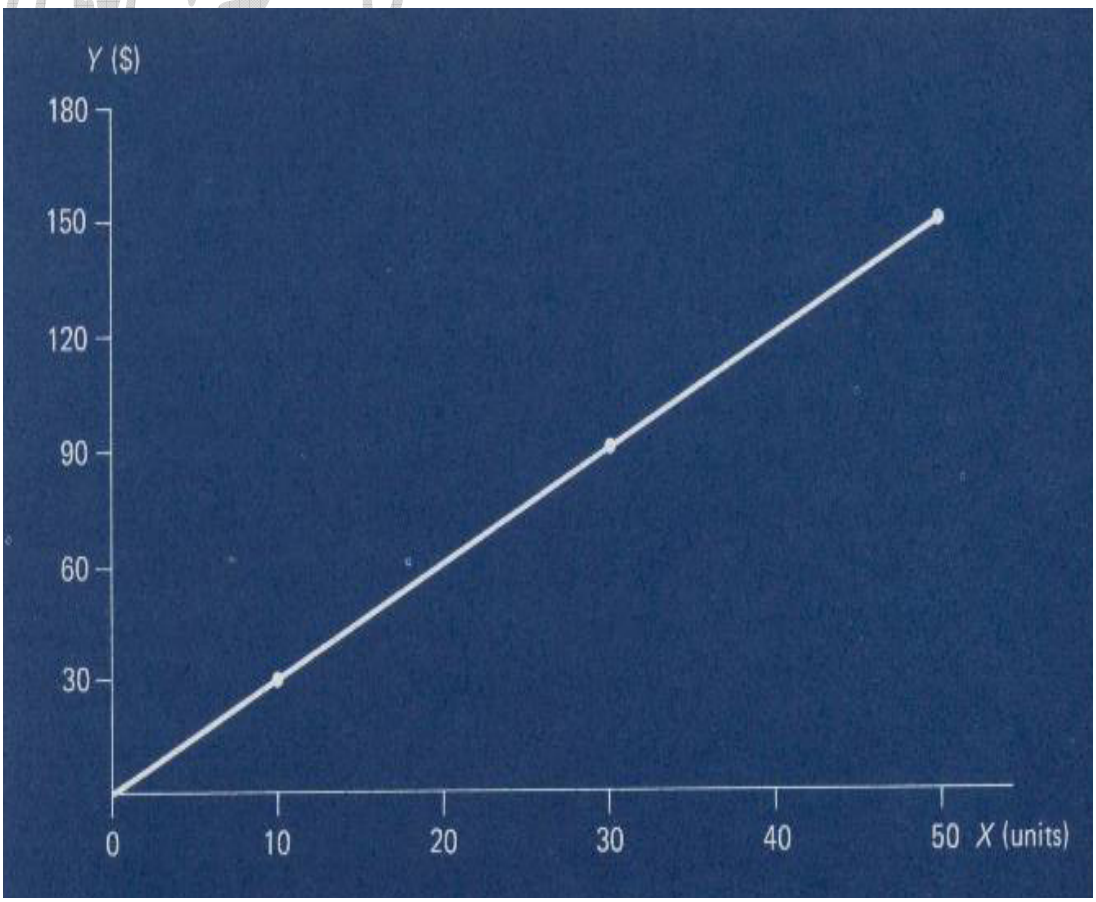
7.2 Relationships between Variables

The concept of *a **functional relationship*** between two variables is familiar to us. If a functional relationship exists between two variables, then it is possible to represent the relationship by a formula $Y = f(X)$, where X is the independent (or predictor) variable, and Y is the dependent (or predicted) variable.

Example 7.1

Suppose for every unit of a product sold, a company makes a profit of \$3. Let X = number of units sold, and Y =total profit. Then. $Y = 3X$. Illustrate this linear relationship.

Figure 7.1 Graph of the functional relation Y= 3X



Solution:

This linear functional relationship is illustrated in Figure 7.1.
For example, if X = 10, Y= 3(10) =30; if X = 30, Y = 3(30) =90 and if X = 50; Y = 3(50) =$150, Notice that all three of the pairs (X, Y) of points fall exactly on a straight line.

In Example 7.1, the functional relationship is linear. An example of a nonlinear functional relationship is Y= Xz. If, for example, X= 2, then Y= 4. A graph of the functional relationship Y = Xz is illustrated in figure 7.2.

In a statistical relationship, the variables are not perfectly related as they are in a functional relationship. The pairs of points (X, Y) will not all lie perfectly on the curve representing the relationship between the variables.

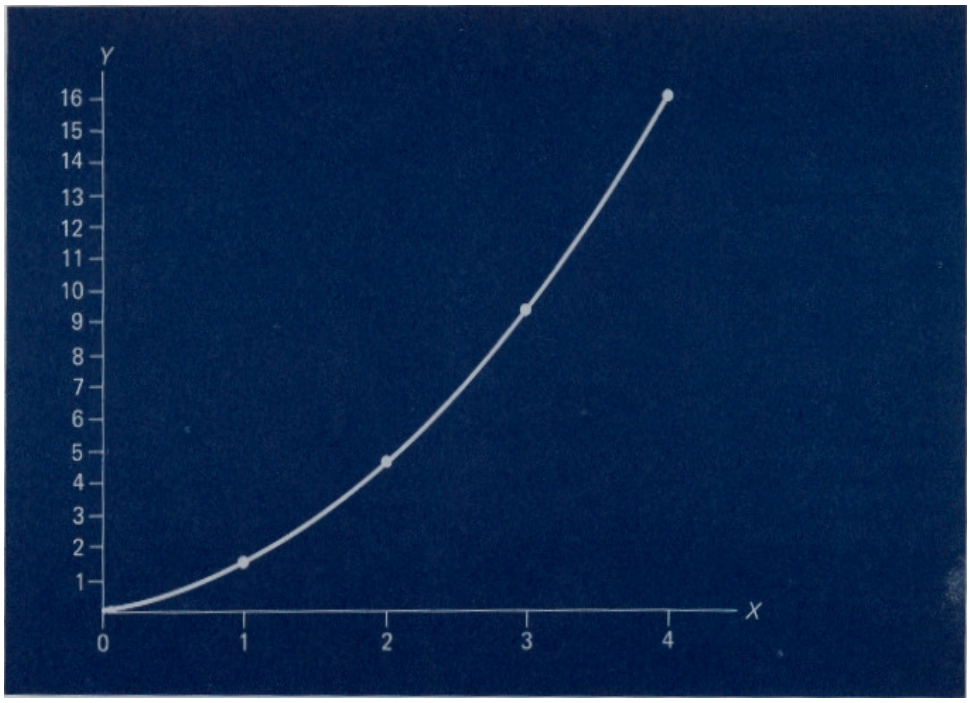
Example 7.2

An example of a statistical relationship is the relationship between heights and weights of adult. males. Table 7.1 contains the heights and weights of ten randomly selected males. Plot this data as a graph similar to Figure 7.1.

Solution:

These data are plotted in figure 7.2. Clearly, the taller a man is, the more he weighs. But, the relationship is not a perfect one, as is evident in Figure 7.3. The line in figure 7.2 has been drawn to fit reasonably well through the ten points, and the points are scattered about this line. The

Figure 7.2 Graph of the nonlinear functional relation Y= Xz

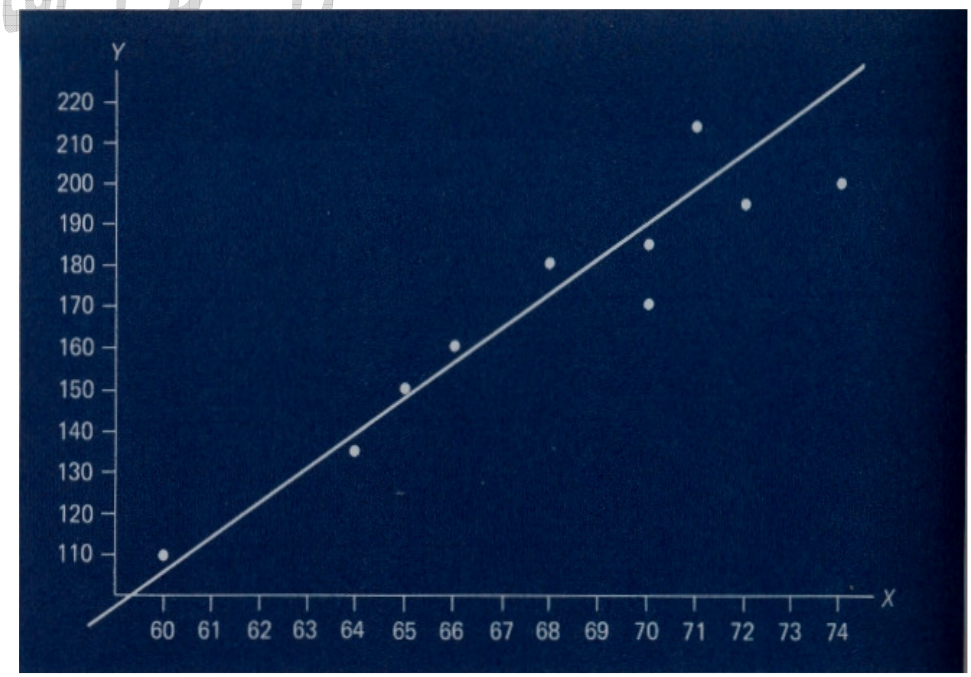


scattering of points suggests that some of the variation in weight is not accounted for by height alone. For instance, two men are 70 inches tall, but their weights differ- 185 pounds and 170 pounds. The variation in weight not accounted for by height alone may be considered to be random in nature, but may also be due to the failure to include other important independent predictor variables. The randomness of the scattering of points about the fitted line is an important element in assessing the validity of a regression model. Before explaining how we fit a line to the data, we must first describe the regression model and the assumptions necessary for its correct application.

Table 7.1 Heights (X) and weights (Y) of ten randomly selected adult males

Height (X) inches	Weight (Y) pounds
60	110
65	150
74	200
70	185
70	170
66	160
68	180
72	195
64	135
71	215

Figure 7.3 Plot of data in Table 7.1



7.3 Simple Linear Regression Model

The simple linear regression model is a mathematical way of stating the statistical relationship that exists between two variables. The two ***principle elements*** of a statistical relationship are:

(1) the tendency of the dependent variable Y to vary in a systematic way with the independent variable X, and

(2) the scattering of points about the "curve" that represents the relationship between X and Y.

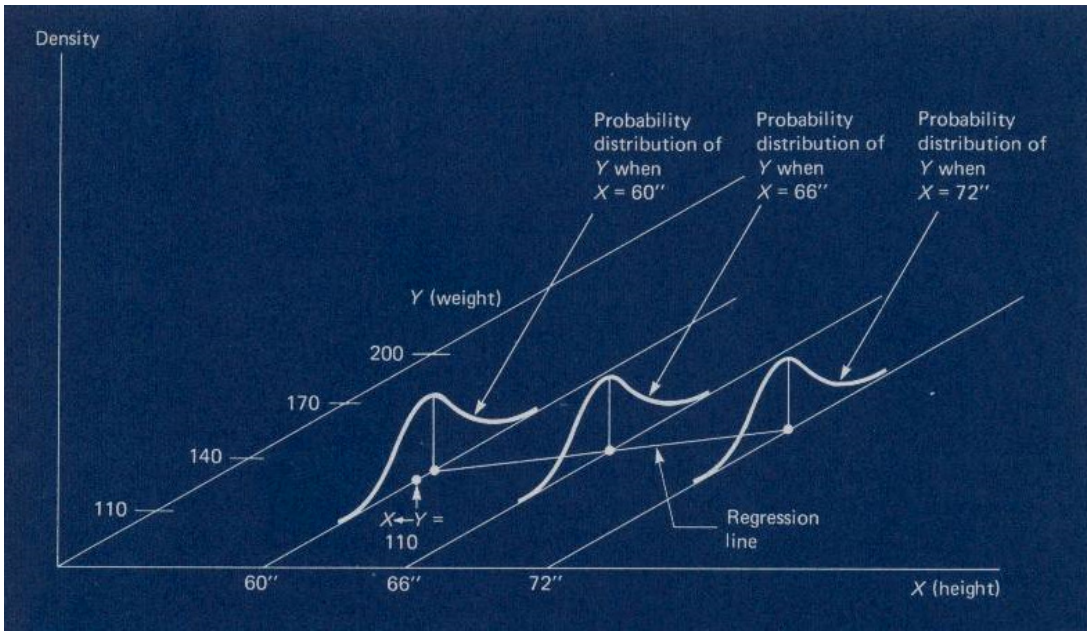
These two elements of a statistical relationship are represented in a simple linear regression model by ***assuming that***:

(i) there is a probability distribution of Y for each value of X, and

(ii) the means of these probability distributions fall perfectly on a line.

These two assumptions are illustrated in figure 7.4 for the Example 7.2 data. The systematic way in which Y varies as a function of X is identified as a straight line, the regression line of Y on X. The regression line goes perfectly through the means of the conditional probability distributions of Y, given a value of X. The data are collected by taking random samples from the conditional probability distribution of Y for values of X. For example, from Table 7.1, when X = 60 inches, Y was observed to be 110 pounds. This particular value of Y represents a random sample of size one drawn from the conditional probability distribution of Y when X = 60 inches.

Figure 7.4 Graphical form of the simple linear regression model



The formal statement of the simple linear regression model is:

$$Y_i=\beta_0 +\beta_i.X_i+e_i, \qquad i= 1,2,...,n,$$

where:

Y_i = Value of the dependent variable in the ith trial ,
β_0, β_1= Parameters in the model,
X_i = Value of the independent variable in the ith trial,
e_i = Random error term in the ith trial.

By trial, we mean an observed value of Y for a fixed value of X. For example, in Table 7.1, the data are generated by making ten trials of the simple experiment. In the first trial, we set X = 60 inches and from the distribution of weights for all. adult males who are 60 inches tall, we sample one whose weight is 110 pounds.

There are actually two ways we may acquire the needed sample information as given in Table 7.1: by experimentation or by survey. To experimentally generate the sample data, we would select a set of values of X, and for each we would randomly sample one or more values of Y. For example, we may be interested in the yield of a chemical compound Y measured in grams as a function of pressure X in a chemical production process.

We could select a set of pressures (values of X) and then run the production process at each pressure setting one or more times to produce observations on Y. Alternatively, we could generate the sample data by taking a survey. For example, we could randomly sample ten adult. males to determine their heights and weights. But, the survey method has the disadvantage that we must take whatever values of X(height) occur in the survey; the selection of the set of values of X, the independent variable, is out of our control. We might be so unfortunate, for instance, to find that all ten men in our survey were 64 inches tall. We ideally want a spread of X values over the range of interest and over which the regression line will be built.

It is always better to produce the sample data by experimentation, if possible, for then we can control the independent variable X-the experiment can be designed to suit our needs. When experimentation is not possible, surveys must be used to generate the data.
The ***assumptions*** corresponding to the simple linear regression model are:

1. For the ith trial, the expected value of the error component e_i is zero [$E(e_i) = 0$], and the variance of the error component [$V(e_i)$] is σ^2 and is constant for all values of i, i= 1, 2, ..., n.
2. The error components in any pair of trials, say the ith and the jth, are uncorrelated.
3. The terms β_o and β_1 in the model are parameters whose values are typically unknown and must, therefore, be estimated from sample data. Further, X; is considered to be a known constant in the model.

The ***consequences*** of these ***assumptions*** are:

1. The observed value of Y in the i th trial, Y_i, is the sum of two components; a constant and a random variable:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

↑ constant ↑ random variable

2. By using the expectation operator rule given in Chapter 5, we get

$E[Y_i]=E[\beta_0+ \beta_1X_i+e_i]= \beta_0+ \beta_1X_i+E[e_i]$. But,

$E[e_i]=0$, so $E[Y_i]= \beta_0+ \beta_1X_i$.

Thus, the mean of the conditional probability distribution of Y given a value of X, denoted by $\mu_{y/x}$, is equal to $\beta_0 + \beta_1 X_i$.
And. therefore, the regression function corresponding to the regression model is

$$E[Y]= \beta_0 + \beta_1 X.$$

3. By using the variance operator rule given in Chapter 5,

$$V[Y_i]=V[\beta_0+ \beta_i\ X_i+e_i]=V[\beta_0+ \beta_1 X_i]+V[e_i]$$
$$=0+V[e_i].$$

But, $V[e_i] = \sigma^2$, so $V[Y_i] = \sigma^2$.

Thus, the variance of the conditional probability distribution of Y given a value of X, denoted by $\sigma^2_{y/x}$, is equal to σ^2 and each conditional probability distribution has the same variance, σ^2.

4. The observed value of Y in the ith trial is larger or smaller than $\mu_{y/x}$ by the amount e_i, the value of the error component in the ith trial.

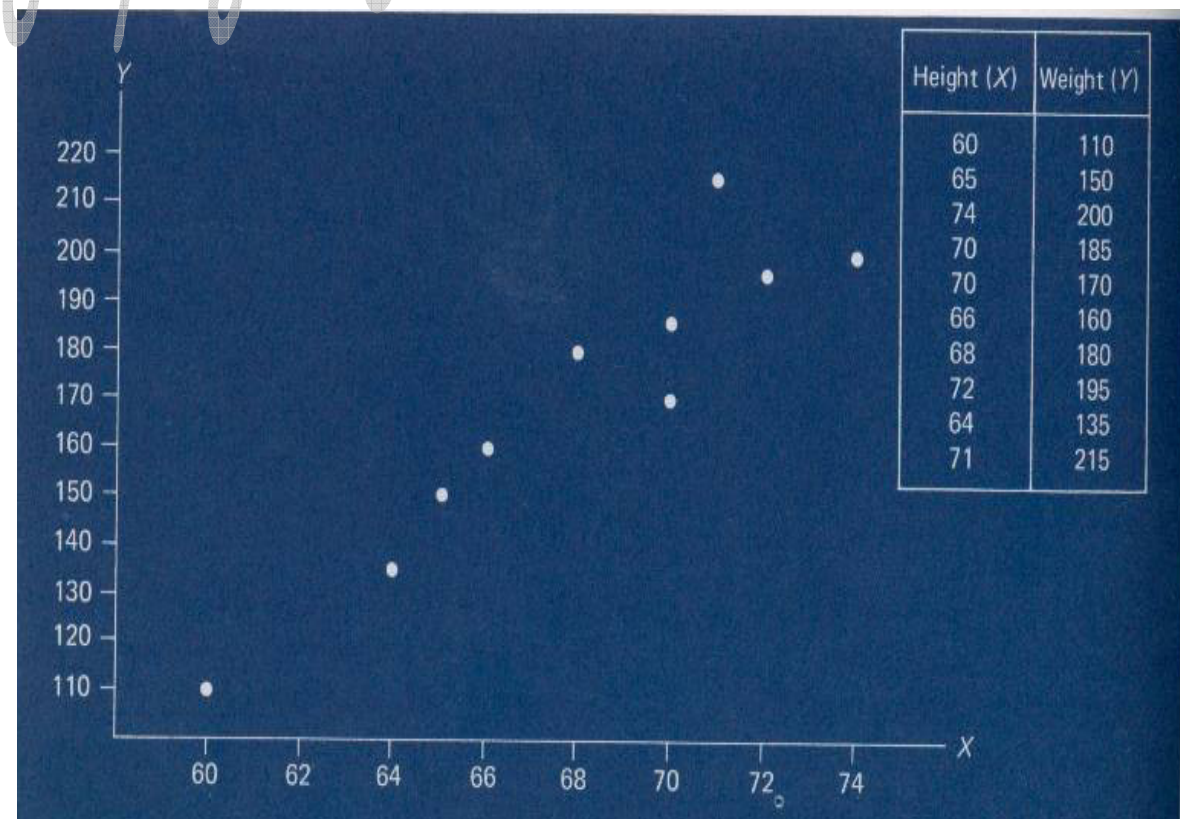
5. By the second assumption , the outcome in any trial is not affected by or does not itself affect the error term in any other trial .

7.4 Fitting of a Simple Linear Regression Model

Since β_0 and β_1, are generally not known in a regression problem, they must be estimated from sample data taken on the dependent variable Y for a number of values of the independent variable X. These pairs of sample values are obtained either by experimentation or by survey. The data given in Table 7.1 were determined by survey- 10 adult. males were selected at random and their heights and weights were recorded.

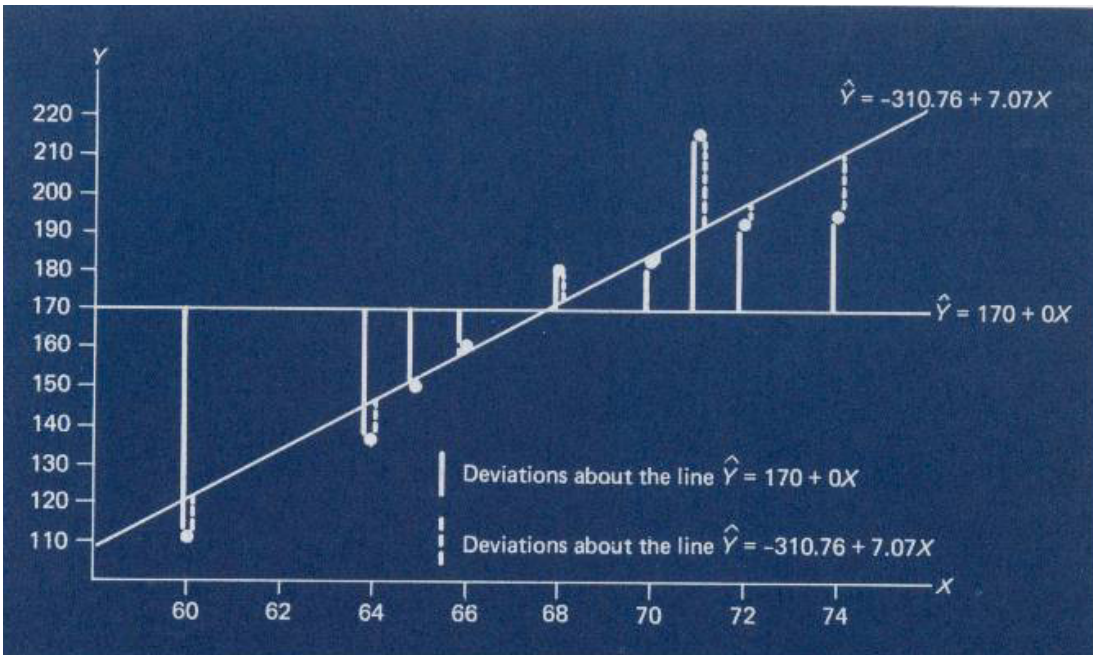
Returning to the data given in Table 7.1, we will first produce a scatter plot of these data. The scatter plot is given in figure 7.5. In figure 7.6, we have superimposed two "fitted" lines through this scatter of points,

Figure 7.5 Scatter plot of data given in table 7.1



Height (X)	Weight (Y)
60	110
65	150
74	200
70	185
70	170
66	160
68	180
72	195
64	135
71	215

Figure 7.6 Scatter plot and fitted lines for data in Table 7.1



$\hat{Y} = 170 + 0\,X$ and $\hat{Y} = -310.76 + 7.07X$, respectively. It is apparent in Figure 7.6 that the line $\hat{Y} = -310.76 + 7.07X$ fits the given data "better," but we must establish a criterion to evaluate when one line is "better" than another so that we may find the best fitting line. The criterion we shall use is called least squares. For each sample observation (X_i, Y_i), the least squares criterion considers the deviation of Y_i from its expected value:

$$[Y_i - E(Y_i)] = [Y_i - (\beta_0 + \beta_1 X_i)] = e_i$$

and requires that values of β_0 and β_1, be found which minimize:

$$\text{LS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} e_i^2$$

The specific values of β_0 and β_1, that minimize LS are the regression coefficient estimates, denoted by b_0 and b_1 respectively.

Thus, the least squares criterion requires that we find a line, denoted by $\hat{Y} = b_0 + b_1 X$, such that the sum of the squared vertical deviations between the line and the scatter of points is minimized. In figure 7.6, the vertical deviations corresponding to the line $\hat{Y} = 170 + 0X$, where $b_0 = 170$ and $b_1 = 0$, are indicated. Obviously, the line $\hat{Y} = -310.76 + 7.07X$ in figure 13.11, where $b_0 = -310.76$ and $b_1 = 7.07$, does much better in the least squares sense because its vertical deviations from the scatter of points, when squared and summed, will be less than the sum of squared deviations for the line $\hat{Y} = 170 + 0\,X$.

It turns out that the values of b_0 and b_1 , which minimize LS are solutions to the following two simultaneous equations, which are referred to as the normal equations:

$$\Sigma Y_i = nb_0 + b_1 \Sigma X_i$$
$$\Sigma X_i Y_i = b_0 \Sigma X_i + b_1 \Sigma X_i^2$$

Solving the normal equations for b_0 and b_1, produces the point estimators of β_0 and β_1 respectively. The resulting formulas for b_0 and b_1, are given below:

$$b_1 = \frac{\Sigma X_i Y_i - \dfrac{(\Sigma X_i)(\Sigma Y_i)}{n}}{\Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{n}}$$

$$b_0 = \overline{Y} - b_1 \overline{X}, \quad \text{where} \quad \overline{Y} = \frac{\Sigma Y_i}{n}, \ \overline{X} = \frac{\Sigma X_i}{n}$$

Example 7.3

Let us now fit a simple Linear regression model to the data on heights and

weights given in Table 7.1

Solution:

The computations for determining b_0 and b_1, are given in Table 7.2. The format in this table provides a convenient worksheet for finding the necessary components in the formulas for b_0 and b_1

$$b_1 = \frac{\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{n}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{116{,}745 - \frac{(680)(1{,}700)}{10}}{46{,}402 - \frac{(680)^2}{10}} = 7.07$$

$$b_0 = \overline{Y} - b_1\overline{X} = \frac{1{,}700}{10} - (7.07)\left(\frac{680}{10}\right) = -310.76$$

Thus, the fitted regression line is $\hat{Y}$=- 310.76 + 7.07X and this is the best fitting line based upon the least squares criterion.

TABLE 7.2 Computation worksheet for determining b_0 and b_1 in Example 7.3

Observation	Height (X)	Weight (Y)	X^2	Y^2	XY
1	60	110	3,600	12,100	6,600
2	65	150	4,225	22,500	9,750
3	74	200	5,476	40,000	14,800
4	70	185	4,900	34,225	12,950
5	70	170	4,900	28,900	11,900
6	66	160	4,356	25,600	10,560
7	68	180	4,624	32,400	12,240
8	72	195	5,184	38,025	14,040
9	64	135	4,096	18,225	8,640
10	71	215	5,041	46,225	15,265
Totals	680	1,700	46,402	298,200	116,745

Tutorial 7

I. Distinguish between dependent and independent variables in a regression model.

2. Why is it important to plot a scatter diagram of the relationship between variables in a simple linear regression model?

3. What is meant by "least squares" in a simple regression model?

4. Describe the normal equations and how they are derived.

5. Discuss the assumptions made in using simple linear regression about the distributions of the conditional mean values.

6. What is meant by the coefficient of determination?

7. Plot each of the following sets of data as a scatter diagram. Which curves seem to fit the data best?
   Determine the regression equation for each set of data.

A

Y Vehicle registrations	*X* Miles of primary highway
125	5,022
155	9,984
130	14,738
202	19,921
194	25,021
241	30,550
310	34,729
397	41,001
570	45,143
656	50,002

In Nittany Valley, 1930–70.

B

Y Sales of gasoline (1,000 gallons)	*X* Price of gasoline (cents per gallon)
50.5	37
57.3	36
60.0	35
60.1	31
67.8	29
70.1	27
68.2	26
74.8	25
75.1	23

In Nittany Valley, 1930–70.

C

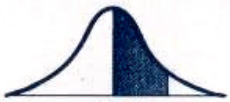
Y Batting average	*X* Age
0.304	21
0.299	24
0.293	27
0.288	28
0.280	30
0.267	32
0.260	33
0.252	34
0.264	35

Nine individuals, selected at random, 1977 "Happy Valley" softball league.

Table 1 Binomial Distribution Probability

		p									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	2	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8755	0.8400	0.7975	0.7500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7182	0.6480	0.5748	0.5000
	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3	1.0000	0.9995	0.0078	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
	3	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6562
	4	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0 0039
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980

Table 2 Standard normal distribution areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
4.0	.4999683									

References

1.William G. Cochran.
Sampling Techniques.

2. Irwin Miller & John E. Freund
Probability and Statistics for Engineers.

3.John E. Freund
Modern Elementary Statistics .

4.Roger C. Pfaffenberger & James H. Patterson
Statistical Methods for Business and Economics .

5.Snedecor , G.W. and Cochran , W.G.
Statistical Methods .